PHYSEDITWORLD: A DATASET FOR PHYSICS-EDITABLE GAME WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Game world models can generate visually plausible gameplay, yet controlling how motion responds to changes in physical parameters remains challenging. Studying this capability requires comparable observations in which a physical parameter varies while the scene and interaction remain fixed. We introduce PhysEditWorld, a multimodal dataset for physics-editable game world modeling, together with our open-source UE5 DataFactory plugin. The current release focuses on gravity. The plugin replays recorded interactions under different gravity settings while preserving the scene, initial state, action trace, controller configuration, and camera policy. This design produces matched motion sequences with explicit gravity labels and synchronized visual, action, camera, and engine-state annotations. PhysEditWorld spans 12 authored UE5 scenes, 100+ hours of gameplay interactions, and 60M+ rendered frames. Fine-tuning on PhysEditWorld strengthens gravity-dependent motion responses in video generation and improves Qwen3-VL-8B gravity-class accuracy from 24.71% to 95.29% on held-out scenes. The publicly released dataset and plugin support controlled data collection, training, and evaluation of game world models under explicit physics edits.

1 INTRODUCTION

Generative game world models increasingly support visual prediction, user interaction, and persistent environments (Bruce et al., 2024; Alonso et al., 2024; Valevski et al., 2025; Che et al., 2025; Robbyant Team et al., 2026; Wang et al., 2026). However, reproducing motion seen during training does not establish that a model responds to an edited physical parameter. In game authoring, changing gravity should alter a jump or fall while retaining the scene and the player's input. A visually convincing clip can still fail this requirement by changing the action, relying on a camera cue, or producing nearly identical motion across gravity settings.

We study *physics-editable game world modeling*: conditioning a rollout on an explicit physical parameter and evaluating the resulting change under matched initial and interaction conditions. Our current study focuses on gravity. Gravity is a useful starting point because its effects are visible in airtime, fall progression, jump arcs, and contact timing. A single video cannot isolate these effects from different actions, spawn states, or camera controls. The required data unit is therefore a matched replay group in which one recorded interaction is expanded under multiple gravity settings.

PhysEditWorld provides this structure through a UE5 replay-and-rendering pipeline. Each matched group shares a scene, initial state, semantic action sequence, controller configuration, and camera policy. Changing gravity can change the resulting character and camera trajectories; it does not change the prescribed controls or camera policy. The data include 12 authored scenes and 60M+ rendered frames, with action traces and engine logs linking visual motion to the simulated intervention. The dataset is publicly released, with the release scope and asset restrictions stated separately from the evaluation protocol.

We examine dataset utility by comparing the same backbone before and after PhysEditWorld fine-tuning. The scene-disjoint VLM evaluation measures gravity prediction, while generation experiments separately report target-event success, pair coverage, and gravity ordering; the rooftop input is retained as a qualitative appearance-transfer case. These comparisons measure the practical benefit of additional dataset supervision. They do not, by themselves, isolate the contribution of replay matching from the other benefits of fine-tuning. Our contribution is the gravity-intervention data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

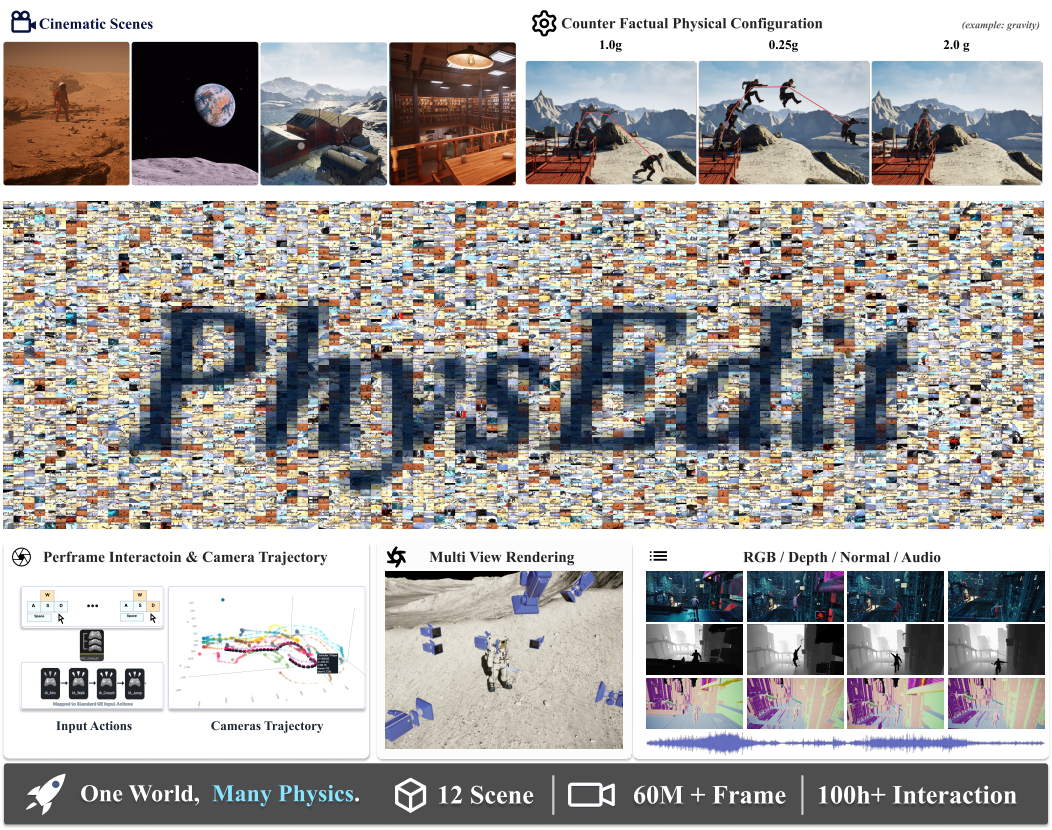

Figure 1: PhysEditWorld records gameplay once and replays it under different gravity configurations. The pipeline supports eight synchronized camera views and aligned visual, action, camera, and state annotations. The evaluated intervention in this paper is gravity.

resource and its reproducible task formulation, rather than a new generative architecture or a claim of universal physical understanding.

Specifically, we contribute (i) a multimodal dataset organized around matched gravity interventions in gameplay; (ii) our open-source UE5 DataFactory plugin and replay-and-rendering pipeline, which separate interaction capture from gravity variation and synchronized export; and (iii) utility studies and an evaluation design that distinguish event response, pair coverage, gravity ordering, and metric calibration. This separation prevents a model that merely produces faster motion from being treated as a quantitatively accurate simulator.

2 Related Work

Game world models. Game world models have evolved from single-game neural simulators to open-domain interactive video generators. Early works such as GameGAN, Playable Video Generation, Playable Environments, and Promptable Game Models studied controllable neural simulation and interactive video manipulation from gameplay or video data (Kim et al., 2020; Menapace et al., 2021; 2022; 2024; Xu et al., 2026). Recent systems, including Genie, Oasis, DIAMOND, GameNGen, GameGen-X, GameFactory, MineWorld, YUME, LingBot-World, Matrix-Game 3.0, and related interactive video world models, scale action-conditioned generation, open-world exploration, long-horizon consistency, ID consistency, and real-time interaction (Bruce et al., 2024; Decart et al., 2024; Alonso et al., 2024; Valevski et al., 2025; Che et al., 2025; Yu et al., 2025; Guo et al., 2025; Mao et al., 2025; Robbyant Team et al., 2026; Wang et al., 2026; Chen et al., 2025; Hu et al., 2026). While these models learn rich visual and interaction dynamics, physical rules are usually absorbed as implicit dataset regularities rather than exposed as editable variables. PhysEditWorld addresses

this gap with explicit gravity annotations and matched replays under fixed scene, action, controller, and camera conditions.

Video and world-model datasets. Existing game and world-model datasets mainly emphasize scale, action control, exploration coverage, or state supervision. MineRL, VPT, and MineDojo provide Minecraft demonstrations, video pretraining data, simulator environments, or task suites for embodied agent learning (Guss et al., 2019; Baker et al., 2022; Fan et al., 2022). OGameData/GameGen-X and GF-Minecraft/GameFactory target open-world or action-controllable game video generation (Che et al., 2025; Yu et al., 2025); Sekai and YUME focus on world exploration and interactive generation (Li et al., 2025c; Mao et al., 2025); and WildWorld and MultiWorld add explicit state, multi-agent, or multi-view supervision (Li et al., 2026; Wu et al., 2026). Related resources also study high-DoF action-to-video control or physics-related gameplay failures (Wang et al., 2025c; Taesiri et al., 2022). However, these datasets are not organized around matched physical interventions. PhysEditWorld instead replays the same authored scene, interaction trace, character controller, and camera policy under different gravity configurations, making the physical edit directly comparable.

Physics benchmarks and controllable generation. Many benchmarks study physical understanding from visual data. ShapeStacks and ADEPT probe object stability and expectation violation, while PHYRE, I-PHYRE, IntPhys, IntPhys2, CLEVRER, CATER, and Physion evaluate intuitive physics, compositional actions, causal reasoning, intervention, and future prediction (Groth et al., 2018; Smith et al., 2019; Bakhtin et al., 2019; Li et al., 2024b; Riochet et al., 2022; Bordes et al., 2025; Yi et al., 2020; Girdhar & Ramanan, 2020; Bear et al., 2021). CoPhy, ComPhy, CRIPP-VQA, ContPhy, CausalVQA, and QuantiPhy further study counterfactual dynamics, hidden physical properties, causal alternatives, and quantitative physical quantities (Baradel et al., 2020; Chen et al., 2022; Patel et al., 2022; Zheng et al., 2024; Foss et al., 2025; Li et al., 2025b). Video-generation benchmarks such as VBench, VBench-2.0, VideoPhy, PhyGenBench, Physics-IQ, PhysInOne, WorldScore, and NewtonRewards evaluate physical plausibility, intrinsic faithfulness, world-generation quality, or Newtonian motion consistency (Huang et al., 2024; Zheng et al., 2025; Bansal et al., 2025; Meng et al., 2025; Motamed et al., 2026; Zhou et al., 2026; Duan et al., 2025; Le et al., 2025; Zhang et al., 2026b;a). Recent controllable generation methods condition on force, torque, force fields, physical parameters, Newtonian dynamics, or learned physical priors (Liu et al., 2024; Gillman et al., 2025; Wang et al., 2025a; Zhang et al., 2025; Yuan et al., 2026; Narayanan et al., 2026; Li et al., 2025a). These works provide important reasoning, evaluation, and generation tools, but they do not provide interactive game rollouts where one physical rule is edited while scene content and interaction are held fixed. We therefore frame the current contribution around the matched gravity-game setting rather than claiming novelty for physical intervention or counterfactual simulation in general.

Synthetic simulation platforms. Simulation platforms enable controllable data generation and ground-truth annotation. ALE, Procgen, MineRL, and CARLA support standardized environments for games, reinforcement learning, and driving (Bellemare et al., 2013; Cobbe et al., 2020; Guss et al., 2019; Dosovitskiy et al., 2017). AI2-THOR, RoboTHOR, ProcTHOR, Habitat, Gibson, iGibson, TDW, Kubric, VirtualHome, and BEHAVIOR-1K provide controllable 3D simulation with rich sensor outputs, physical interaction, programmatic actions, or scalable embodied-AI environments (Kolve et al., 2017; Deitke et al., 2020; 2022; Savva et al., 2019; Xia et al., 2018; Li et al., 2022; Gan et al., 2021; Greff et al., 2022; Puig et al., 2018; Huang et al., 2025; Li et al., 2024a). Robotics simulators such as SAPIEN, RLBench, ManiSkill2, and Isaac Gym support articulated-object interaction, manipulation, and high-throughput physics simulation (Xiang et al., 2020; James et al., 2020; Gu et al., 2023; Makoviychuk et al., 2021). UnrealCV connects Unreal Engine to computer-vision pipelines, while UnrealZoo scales UE environments for embodied AI (Qiu & Yuille, 2016; Zhong et al., 2025). PhysEditWorld builds on UE but targets a different protocol: matched replay under edited gravity.

The table summarizes the intended comparison axes; entries describe the released resource interfaces and do not imply that every resource is designed for the same task. PhysEditWorld's distinguishing axis is matched gravity intervention with synchronized gameplay, rather than raw modality count alone.

Table 1: Comparison with related datasets and data-generation resources. RGB-D-N-A denotes synchronized RGB, depth, normal, and audio. Edit phys. denotes whether physical attributes are explicitly editable; PhysEditWorld focuses on gravity in the current release. Camera ann. denotes camera parameters or trajectories. Action/control denotes recorded actions, controls, or interaction traces. ✓: yes, ✗: no, △: partial.

Resource	Multi-view	RGB-D-N-A	Edit phys.	Camera ann.	Action
ALE (Bellemare et al., 2013)	✗	✗	✗	✗	✓
Procgen (Cobbe et al., 2020)	✗	✗	✗	✗	✓
MineRL (Guss et al., 2019)	✗	✗	✗	△	✓
Sekai (Li et al., 2025c)	△	△	✗	✓	✗
GameFactory / GF-Minecraft (Yu et al., 2025)	✗	✗	✗	✗	✓
OGameData / GameGen-X (Che et al., 2025)	△	✗	✗	✓	△
CLEVRER (Yi et al., 2020)	✗	✗	✗	✗	✗
Physion (Bear et al., 2021)	✗	△	✗	△	✗
VBench (Huang et al., 2024)	✗	✗	✗	✗	✗
PhyGenBench (Meng et al., 2025)	✗	✗	✗	✗	✗
VideoPhy (Bansal et al., 2025)	✗	✗	✗	✗	✗
Physics-IQ (Motamed et al., 2026)	✗	✗	✗	✗	✗
PhysInOne (Zhou et al., 2026)	✓	△	✗	△	✗
PhysEditWorld	✓	✓	✓	✓	✓

3 PhysEditWorld Dataset

The atomic data unit is a *matched replay group*. Its members share the authored scene, initial state, action trace, controller configuration, and camera policy, while gravity is deliberately varied. An individual rollout corresponds to one gravity configuration and contains synchronized camera streams, action traces, camera parameters, and engine states. RGB observations can be accompanied by depth, surface normals, object masks, audio where available, and visual captions. Auxiliary-modality availability is recorded per sample; pipeline support should not be interpreted as complete modality coverage for every released sample. Appendix E describes the layout. Split and evaluation accounting is taken from the versioned release manifest rather than reconstructed from prose.

PhysEditWorld contains 12 UE5 scenes, 100+ hours of gameplay interactions, and 60M+ rendered rollout frames. Capture uses 1280×720 resolution at 30 FPS and supports eight camera views. The authored environments include indoor, outdoor, urban, planetary, sports, and object-interaction settings. The quantitative studies use a subset of the corpus, described in Section 5.1; corpus size and experimental sample count are not interchangeable.

Table 1 preserves the comparison with prior game, world-model, and physics-oriented resources. These resources provide complementary forms of visual scale, action control, camera information, physical reasoning, or simulation infrastructure, but generally do not combine synchronized multi-view gameplay, explicit gravity editing, and matched replay under a fixed scene and action trace. The table describes resource interfaces rather than claiming that every prior dataset targets the same task.

4 Replay-and-Rendering Pipeline

The pipeline operates directly on artist-ready UE5 levels through our open-source in-editor DataFactory plugin, integrating data collection into the game-development workflow. Scene preparation, semantic input recording, controlled replay, and synchronized export form four connected stages (Figure 2). Separating interaction capture from replay expansion allows a validated action sequence to be reused under multiple gravity settings without recollecting the player's behavior. The plugin and the accompanying replay/export interface are part of the contribution, rather than hidden collection infrastructure.

Scenario preparation and input capture. The open-source DataFactory plugin registers the level, interactable assets, spawn state, controller, collision setup, and camera anchors. Human play, scripted sequences, or navigation agents provide semantic actions through UE5's Enhanced Input

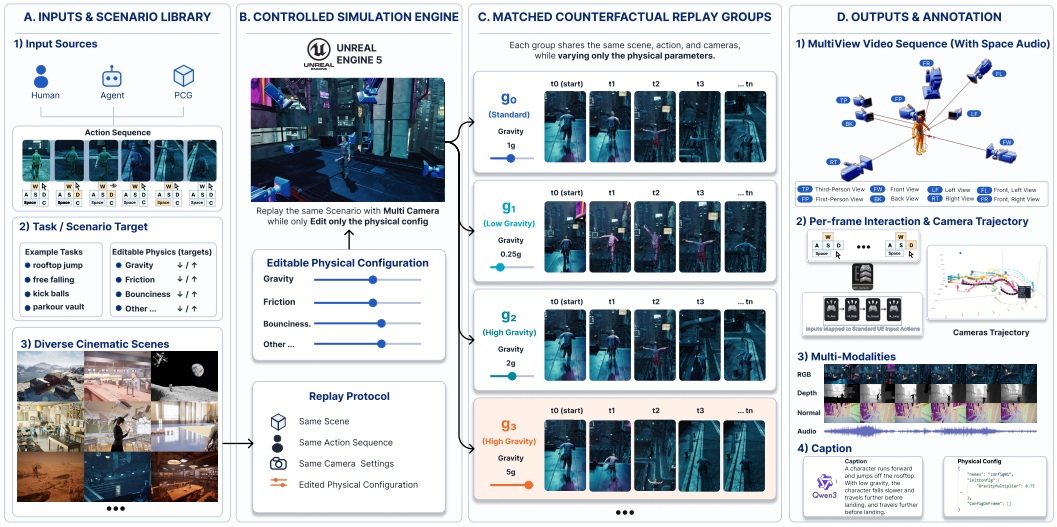

Figure 2: The pipeline prepares native UE5 scenarios, records semantic inputs, replays matched groups, and exports synchronized observations and metadata. Friction and restitution controls shown in the interface illustrate extension points; all quantitative claims here concern gravity.

System. Recorded inputs include movement axes, jump commands, camera deltas, and button states. Replaying these actions through the same gameplay stack preserves the control sequence while allowing the resulting motion to change.

Scenario preparation emphasizes physical visibility and replay stability. Artists select jumps, falls, object interactions, and height-varying traversal, then validate the associated assets, spawn points, collision geometry, and camera anchors. Semantic Input Actions preserve the intended control commands while avoiding dependence on raw device events such as keyboard layouts, mouse sensitivity, and platform-specific input APIs. This representation fixes the requested interaction, while the character's resulting trajectory remains an outcome of the simulation.

Controlled replay. Let A be the scene, s_0 the initial state, S the action trace, M the controller configuration, and π_c the camera policy. For gravity multiplier g, the intended replay contract is

$$x_g = F(A, s_0, S, M, g, \pi_c), \qquad (A, s_0, S, M, \pi_c) \text{ fixed within each group.} \tag{1}$$

The intervention changes gravity rather than the recorded input or animation configuration. Character animation uses a shared motion-matching and retargeting stack with fixed configuration across gravity variants. State-driven animation can still respond to a changed trajectory. Similarly, a character-mounted camera follows the changed motion under the same attachment and control policy. A fixed camera policy therefore does not imply an identical camera path. Repeated replays at unchanged gravity provide a way to audit residual simulation variability; the audit fields are specified in Appendix A.5.

An action trace is scene-bound because its meaning depends on local geometry and interaction targets. Batch expansion operates on compatible scene–action pairs and physical configurations, rather than assigning every trace to every scene. Applying the same actions at different gravity levels can lead to different collisions or failed landings; these are legitimate outcomes of the intervention.

Synchronized export and quality control. Movie Render Queue and runtime logging export visual streams and metadata on a shared rollout timeline. Camera lag is disabled during capture, and multiple camera components observe the same replay instance. Logs include player and camera transforms, object states, input events, and gravity settings. Qwen3-VL-8B-Instruct generates clip captions from sampled frames without simulator metadata (Bai et al., 2025). Rendering failures, broken synchronization, severe clipping, and unstable simulation are filtered during dataset construction. This production-stage quality control is distinct from evaluating generated videos: a model's failure to move or fall must remain part of the evaluation outcome.

Each exported record is indexed by rollout identity and frame time. Frame-level metadata link observations to character and camera transforms, while event-level metadata retain input and capture events. Physical-configuration files store the gravity multiplier separately from visual captions. This organization supports both video-only model inputs and simulator-referenced analysis without confusing a caption-derived estimate with an engine measurement.

5 Dataset Utility

We evaluate the resource at three levels: gravity-conditioned generation, action-conditioned first-person rollouts, and video-language gravity prediction. The primary comparison uses the same backbone before and after PhysEditWorld fine-tuning. This comparison measures dataset utility. Matched replay controls non-gravity variation in the training data; whether this organization accelerates convergence remains an untested hypothesis and is not a claimed experimental result.

5.1 Setup and Scene-Disjoint Evaluation

Splits. The evaluation manifest contains 1,530 $(\text{scene}, \text{action}, \text{gravity})$ tuples from nine training scenes and 170 tuples from three held-out scenes. Scenes are assigned to a split first. All gravity variants and camera views of a replay group remain in that split, and at most one training clip is sampled per tuple. Thus, the held-out evaluation is cross-scene, while each matched group contains the same scene under different gravity conditions. These two properties address different confounds: scene separation tests transfer to unseen environments, and within-group matching prevents a single scene identity from determining the gravity label.

The training split contains multiple gravity multipliers, while the generation case study additionally evaluates 0.25, 0.75, and 10. These case-study settings are reported as a separate example and are not used to infer aggregate coverage or extrapolation performance.

Models and training. We compare Wan2.2-TI2V-5B before and after LoRA fine-tuning for gravity-conditioned video generation (Wan Team et al., 2025; Hu et al., 2022). For action-conditioned generation, we compare LingBot-World before and after fine-tuning and include frozen Matrix-Game 3.0 as an external reference (Robbyant Team et al., 2026; Wang et al., 2026). The recorded generative training recipe uses rank-128 LoRA on attention projections, five epochs, AdamW with learning rate 10^{-4}, batch size eight, and eight H100 GPUs. Source clips are five seconds at 30 FPS and 1280×720. A prompt field such as `gravity: 1.0g` supplies the gravity condition. Backbone-specific temporal sampling, resolution preprocessing, and inference settings must be reported separately from the source-video format (Table 6). Kling 3.0 and Seedance 2.0 are qualitative references only.

5.2 Gravity Response and Evaluation Metrics

The existing analysis estimates a camera trajectory using VGGT (Wang et al., 2025b), extracts a downward progress signal $q(t)$, differentiates it, and fits

$$v(t) = at + b. \tag{2}$$

Here a is an acceleration proxy and R^2 measures fit quality. VGGT translations are expressed in a learned canonical scale, so neither high R^2 nor a monotonic response alone establishes physical accuracy. Separate monocular reconstructions may have different scale and orientation, even for matched inputs; cross-clip comparisons require calibration to a shared reference. We therefore report event response, pair coverage, and conditional gravity ordering separately. Metric acceleration is reported only when the shared reference has independently known dimensions and the complete measurement pipeline agrees with engine replay. Appendix A specifies the calibration procedure using static geometry, engine-reference validation, and fixed preprocessing.

The extended reporting design retains every scheduled generation in the denominator and uses two different levels of validity. *Event success* is a clip-level outcome: the generated video leaves the support and contains the target motion interval, such as a measurable fall. *Pair coverage* is a comparison-level outcome: a scheduled unequal-gravity pair counts only when both clips have the event and yield a usable, commonly calibrated estimate with a valid interval. Pair coverage is therefore not another

Table 2: Aggregate generation report. Group denominators and bootstrap intervals are reported for the evaluation set. Event: clip-level target-event success; pair coverage: scheduled unequal-gravity pairs with two successful events and usable common calibration; A_{all}: pair-coverage-weighted ordering.

Model	Groups	Event ↑	Pair coverage ↑	A_{valid} ↑	A_{all} ↑
Wan, zero-shot	25	62.4 [58.9,65.8]	58.7 [54.6,62.8]	54.2 [49.0,59.1]	31.8 [28.0,35.7]
Wan + SFT	25	83.6 [79.5,87.1]	76.9 [72.1,81.3]	81.7 [76.8,85.9]	62.8 [57.1,68.1]
LingBot, zero-shot	25	55.8 [51.2,60.4]	51.2 [46.0,56.6]	52.1 [46.7,57.4]	26.7 [22.7,31.0]
LingBot + SFT	25	78.4 [73.8,82.7]	72.3 [67.4,77.0]	76.9 [71.8,81.5]	55.6 [50.1,61.1]
Matrix-Game, frozen	25	49.7 [45.0,54.5]	46.5 [41.0,52.0]	50.8 [44.9,56.3]	23.6 [19.7,27.9]

Table 3: Single-group VGGT case study expanded to six gravity settings. Proxies use the case study's VGGT units and mean R^2 summarizes fit quality; neither quantity is an absolute physical acceleration. Original three-setting values are documented in Appendix A.6; this table does not establish aggregate performance.

Model	a @ $0.1g$	a @ $0.25g$	a @ $0.5g$	a @ $0.75g$	a @ $1g$	a @ $10g$	Mean R^2
Wan2.2-TI2V-5B	0.0003	0.0002	0.0011	0.0046	0.0032	0.0009	0.066
+ PhysEditWorld SFT	0.1057	0.1766	0.2465	0.3983	0.4720	0.6214	0.570

visual-quality score; it measures how often the physical comparison can actually be performed. For a group with scheduled unequal-gravity pairs $\mathcal{P}$, let $m_{ij} = 1$ only when both clips satisfy these conditions. Define

$$C = \frac{\sum_{\mathcal{P}} m_{ij}}{|\mathcal{P}|}, \quad A_{\text{valid}} = \frac{\sum_{\mathcal{P}} m_{ij} \mathbb{1}[(g_i - g_j)(a_i - a_j) > 0]}{\sum_{\mathcal{P}} m_{ij}}, \quad A_{\text{all}} = C A_{\text{valid}}. \tag{3}$$

If no pair is usable, A_{valid} is undefined and $A_{\text{all}} = 0$. The latter counts unavailable pairs as unsuccessful for an end-to-end score; it must be interpreted together with pair coverage and failure reasons, not as a pure physics metric. Chance ordering is 0.5 only for independent, tie-free valid estimates. Aggregation first averages seeds and then replay groups; frames and synchronized views are not independent replicates.

5.3 Gravity-Conditioned Video Generation

The single-group case study shows a stronger gravity response after Wan fine-tuning (Figure 3). Its acceleration proxies at 0.25, 0.75, and 10 are $(0.0002, 0.0046, 0.0009)$ before fine-tuning and $(0.1766, 0.3983, 0.6214)$ afterwards; mean R^2 increases from 0.066 to 0.570. These values are six-setting measurements within one example, not test-set averages. Table 3 makes the six-setting trend explicit while keeping the case-study status visible. The measurements remain in shared VGGT units and are subject to the calibration limits in Appendix A.6; the aggregate evaluation is reported separately in Table 2, with additional breakdowns in Appendix A.7.

5.4 First-Person Generalization Beyond the Training Renderer

Figure 4 evaluates an image-initialized rooftop scenario whose first frame was generated by GPT-Image2, rather than rendered from the training scenes. It is an out-of-distribution (OOD) appearance test relative to the PhysEditWorld fine-tuning corpus; it does not establish absence from the base models' pretraining distributions. Holding the `W` key requests movement over the edge. In this example, the frozen models remain near the platform, while fine-tuned LingBot-World departs and exhibits downward motion that differs between $0.05g$ and $1g$.

This example supplies qualitative evidence of transfer to a new visual input. There is no underlying UE scene or simulator trajectory for this image, so we do not report metric acceleration accuracy for it. We therefore treat it as an appearance-level OOD case rather than a broad generalization benchmark. Broader claims require multiple independent initial images, matched seeds and controls, fixed frame timestamps, and both successful and failed rollouts.

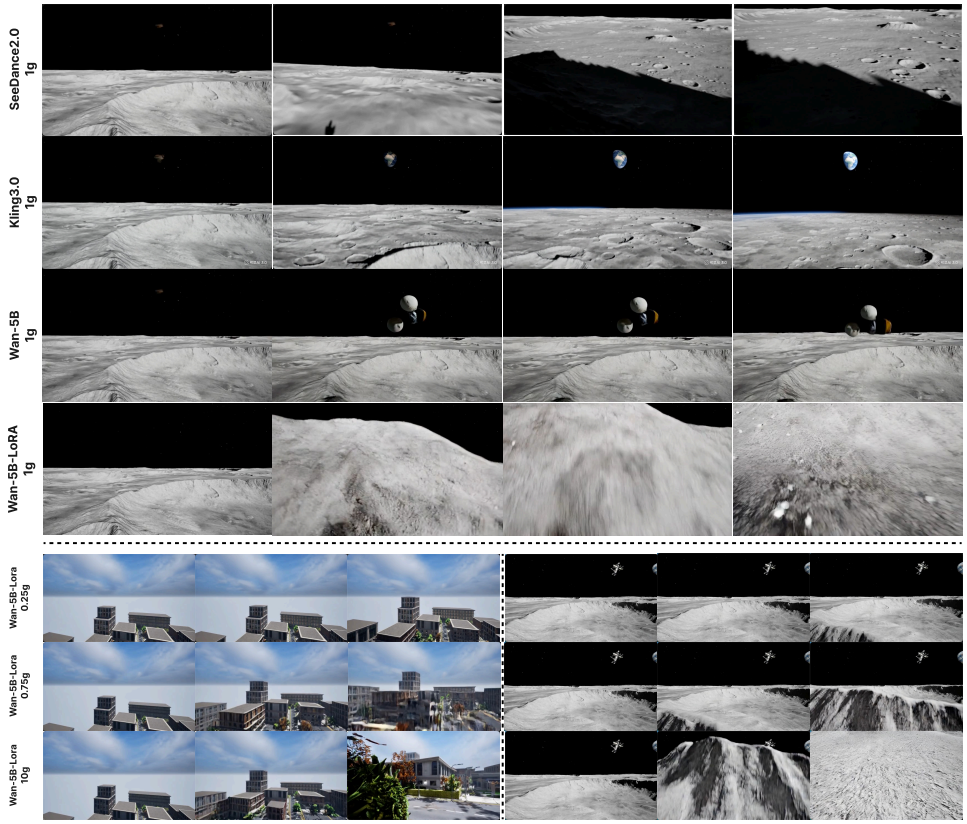

Figure 3: Gravity-conditioned generation examples. Top: Seedance 2.0, Kling 3.0, Wan2.2-TI2V-5B, and fine-tuned Wan at $1g$. Bottom: fine-tuned Wan at $0.25g$, $0.75g$, and $10g$. The displayed examples suggest a gravity-dependent motion response; they do not establish continuous control or dataset-wide accuracy.

Table 4: VLM point estimates on 170 rollouts from three held-out scenes. Uncertainty, scene-level breakdowns, and shortcut diagnostics are reported in the appendix. MAE is in gravity-multiplier units.

Model	Acc. (%) ↑	MAE ↓	Med. APE (%) ↓	Within 10% (%) ↑
Qwen3-VL-8B	24.71	5.4573	95.00	9.48
+ PhysEditWorld SFT	95.29	0.8305	6.22	90.59

5.5 Gravity-Aware Video-Language Understanding

We evaluate Qwen3-VL-8B-Instruct on the 170 held-out rollouts from three unseen scenes (Bai et al., 2025). The model predicts a coarse gravity class and a numerical multiplier. LoRA fine-tuning uses the 1,530 training rollouts; the recorded recipe jitters numerical targets by $\pm 10\%$ and evaluates against the original labels. This jitter is target perturbation, not physical variation in the videos, and does not by itself demonstrate continuous-value generalization. Class boundaries, video sampling, and parsing details are documented in Table 6.

Prediction tasks and metrics. The classification task assigns a clip to low, normal, or high gravity. Numerical prediction reports the multiplier $\hat{g}$ directly. For a positive target g, absolute error is $|\hat{g} - g|$ and absolute percentage error is $100|\hat{g} - g|/g$. We report mean absolute error, median absolute percentage error, and the fraction of predictions satisfying $|\hat{g} - g| \leq 0.1g$. These metrics distinguish coarse gravity recognition from numerical precision; percentage error also prevents large gravity values from solely determining the evaluation.



Figure 4: OOD appearance case study with a GPT-Image2 rooftop input. Frozen Matrix-Game 3.0 and LingBot-World remain near the platform in the displayed runs. Fine-tuned LingBot-World generates departure and differing downward motion. This is a qualitative transfer example, with no engine-ground-truth trajectory.

Table 4 supports improved gravity prediction after dataset fine-tuning: accuracy increases by 70.58 percentage points, and median APE decreases from 95.00% to 6.22%. Training and test scenes are disjoint, and the same scene is represented under different gravity conditions. Consequently, the evaluation does not reuse training-scene identities as test labels, and scene identity alone does not determine gravity within a matched group. Appendix A.4 reports controlled-input diagnostics that complement this split design by examining temporal and foreground evidence. The majority-class baseline contextualizes accuracy under the empirical class distribution.

The two tasks probe complementary uses of physical labels: gravity prediction recovers the edited condition from observed motion, while conditioned generation tests whether specifying that condition changes the output. Matched replay fixes the interaction context across gravity variants, connecting recognition and control within a common experimental setting. Scene-disjoint evaluation further tests transfer beyond the training environments.

6 Conclusion and Future Work

PhysEditWorld provides a dataset and an open-source UE5 DataFactory plugin for studying physics-editable game world models. The pipeline replays the same interaction under different gravity configurations while holding the scene, initial state, action trace, controller configuration, and camera policy fixed. Synchronized visual observations and engine-state annotations make these interventions comparable and support both model training and evaluation. Experiments show improved gravity prediction on held-out scenes and stronger gravity-sensitive motion responses after fine-tuning. Together, the dataset and plugin provide a reusable foundation for collecting controlled gameplay data and studying how world models respond to explicit physical edits.

The current resource focuses on gravity and covers only a limited range of physical properties. It does not yet support fine-grained local physical editing, such as assigning different gravity fields to selected regions or independently modifying physical properties of individual objects. Extending the dataset to friction, restitution, drag, and combined interventions would broaden the range of interactions that can be studied. A further direction is to develop spatially localized and object-specific controls, allowing researchers to examine how local edits affect surrounding motion and interactions. These extensions would move the resource toward richer, more flexible physical control in authored game environments.

REPRODUCIBILITY STATEMENT

The dataset and our open-source UE5 DataFactory plugin are publicly available. Sections 3–4 describe the replay protocol; Appendix E specifies the data schema. Appendix A and `scripts/README.md` document the evaluation protocol and measurement conventions. Dataset access, plugin source, and commercial-asset restrictions are described separately in Appendix B.

AI USE STATEMENT

Qwen3-VL-8B-Instruct produced visual captions, GPT-Image2 produced the rooftop initial image, and an AI writing assistant revised the manuscript and LaTeX layout.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and Francois Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (VPT): Learning to act by watching unlabeled online videos. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.11795.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. PHYRE: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, 2019.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeyppEFvS.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, et al. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025. URL https://arxiv.org/abs/2506.09849.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, 2025.

Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. *arXiv preprint arXiv:2505.21996*, 2025.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. ComPhy: Compositional physical reasoning of objects and events from videos. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PgNEYaIc81Q.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International Conference on Machine Learning*, 2020.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. Project page, 2024. URL https://oasis-model.github.io/.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An open simulation-to-real embodied AI platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2004.06799.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/27c546ab1e4f1d7d638e6a8dfbad9a07-Abstract-Conference.html. Outstanding Paper Award.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Conference on Robot Learning*, 2017.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. MineDojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=rc8o_j8I8PX.

Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T. Kao. CausalVQA: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. ThreeDWorld: A platform for interactive multi-modal physical simulation. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

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*, 2025. URL https://arxiv.org/abs/2505.19386.

Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for compositional actions and TEmporal reasoning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1910.04744.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Daniel Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Oliver Groth, Fabian B. Fuchs, Ingmar Posner, and Andrea Vedaldi. ShapeStacks: Learning vision-based physical intuition for generalised object stacking. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1804.08018.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. ManiSkill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.04659.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: A real-time and open-source interactive world model on Minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

William H. Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. MineRL: A large-scale dataset of Minecraft demonstrations. In *International Joint Conference on Artificial Intelligence*, 2019.

Bin Hu, Zipeng Qi, Guoxi Huang, Zunnan Xu, Ruicheng Zhang, Chongjie Ye, Jun Zhou, Xiu Li, and Jingdong Wang. Identity-consistent video generation under large facial-angle variations, 2026. URL https://arxiv.org/abs/2603.21299.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ting Huang, Zeyu Zhang, Ruicheng Zhang, and Yang Zhao. Dc-scene: Data-centric learning for 3d scene understanding. *arXiv preprint arXiv:2505.15232*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark and learning environment. *IEEE Robotics and Automation Letters*, 5(2): 3019–3026, 2020. URL https://arxiv.org/abs/1909.12271.

Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, and Sanja Fidler. Learning to simulate dynamic environments with GameGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. AI2-THOR: An interactive 3d environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. iGibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 455–465. PMLR, 2022. URL https://proceedings.mlr.press/v164/li22b.html.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. BEHAVIOR-1k: A human-centered, embodied AI benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024a.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning*, 2025a. URL https://arxiv.org/abs/2503.09595.

Puyin Li, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. QuantiPhy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. *arXiv preprint arXiv:2512.19526*, 2025b.

Shiqian Li, Kewen Wu, Chi Zhang, and Yixin Zhu. I-PHYRE: Interactive physical reasoning. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=1bbPQShCT2.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025c.

Zhen Li, Zian Meng, Shuwei Shi, Wenshuo Peng, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. Wildworld: A large-scale dataset for dynamic world modeling with actions and explicit state toward generative ARPG. *arXiv preprint arXiv:2603.23497*, 2026. URL https://arxiv.org/abs/2603.23497.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, 2024. doi:10.1007/978-3-031-73007-8_21. URL https://arxiv.org/abs/2409.18964.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High performance GPU-based physics simulation for robot learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://arxiv.org/abs/2108.10470.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Willi Menapace, Stéphane Lathuilière, Aliaksandr Siarohin, Christian Theobalt, Sergey Tulyakov, Vladislav Golyanik, and Elisa Ricci. Playable environments: Video manipulation in space and time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3584–3593, 2022. URL https://arxiv.org/abs/2203.01914.

Willi Menapace, Aliaksandr Siarohin, Stephane Lathuiliere, Panos Achlioptas, Vladislav Golyanik, Sergey Tulyakov, and Elisa Ricci. Promptable game models: Text-guided game simulation via masked diffusion models. *ACM Transactions on Graphics*, 2024.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *International Conference on Machine Learning*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, March 2026. URL https://arxiv.org/abs/2501.09038.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. Phyco: Learning controllable physical priors for generative motion, 2026. URL https://arxiv.org/abs/2604.28169.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. CRIPP-VQA: Counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9856–9870, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-main.670. URL https://aclanthology.org/2022.emnlp-main.670/.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8494–8502, 2018. URL https://arxiv.org/abs/1806.07011.

Weichao Qiu and Alan Yuille. UnrealCV: Connecting computer vision to unreal engine. *arXiv preprint arXiv:1609.01326*, 2016. URL https://arxiv.org/abs/1609.01326.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Veronique Izard, and Emmanuel Dupoux. IntPhys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2022.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Kevin A. Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth S. Spelke, Joshua B. Tenenbaum, and Tomer D. Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e88f243bf341ded9b4ced444795c3f17-Abstract.html.

Mohammad Reza Taesiri, Finlay Macklon, and Cor-Paul Bezemer. CLIP meets GamePhysics: Towards bug identification in gameplay videos using zero-shot transfer learning. In *2022 IEEE/ACM 19th International Conference on Mining Software Repositories*, pp. 270–281, 2022. doi:10.1145/3524842.3528438. URL https://arxiv.org/abs/2203.11096.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.14837.

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, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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. *arXiv preprint arXiv:2509.20358*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c. URL https://arxiv.org/abs/2508.13104.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Haoyu Wu, Jiwen Yu, Yingtian Zou, and Xihui Liu. Multiworld: Scalable multi-agent multi-view video world models. *arXiv preprint arXiv:2604.18564*, 2026.

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Xia_Gibson_Env_Real-World_CVPR_2018_paper.html.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2003.08515.

Jiawei Xu, Zhizhou Zhong, Zhijian Shu, Mingkai Jia, Mingxiao Li, Jia-Wang Bian, Qian Zhang, Kaicheng Zhang, Jin Xie, Jian Yang, et al. Eponav2: Driving world model with comprehensive future reasoning. *arXiv preprint arXiv:2605.14696*, 2026.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H. Chan. NewtonGen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rJ6N6sunaU.

Haoze Zhang, Tianyu Huang, Zichen Wan, Xiaowei Jin, Hongzhi Zhang, Hui Li, and Wangmeng Zuo. PhysChoreo: Physics-controllable video generation with part-aware semantic grounding. *arXiv preprint arXiv:2511.20562*, 2025.

Ruicheng Zhang, Guangyu Chen, Zunnan Xu, Zihao Liu, Zhizhou Zhong, Mingyang Zhang, Jun Zhou, and Xiu Li. Robostereo: Dual-tower 4d embodied world models for unified policy optimization, 2026a. URL https://arxiv.org/abs/2603.12639.

Ruicheng Zhang, Mingyang Zhang, Jun Zhou, Xiaofan Liu, Zunnan Xu, Zhizhou Zhong, Puxin Yan, Haocheng Luo, and Xiu Li. Mind-v: Hierarchical world model for long-horizon robotic manipulation with rl-based physical alignment, 2026b. URL https://arxiv.org/abs/2512.06628.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. ContPhy: Continuum physical concept learning and reasoning from videos. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61526–61558. PMLR, 2024. URL https://proceedings.mlr.press/v235/zheng24l.html.

Fangwei Zhong, Kui Wu, Churan Wang, Hao Chen, Hai Ci, Zhoujun Li, and Yizhou Wang. UnrealZoo: Enriching photo-realistic virtual worlds for embodied AI. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2412.20977. Highlight.

Siyuan Zhou, Hejun Wang, Hu Cheng, Jinxi Li, Dongsheng Wang, Junwei Jiang, Yixiao Jin, Jiayue Huang, Shiwei Mao, Shangjia Liu, et al. Physinone: Visual physics learning and reasoning in one suite. *arXiv preprint arXiv:2604.09415*, 2026.