WorldExam: Benchmarking World Models from Apparent Appearance to Inherent Reactivity

Yuxue Yang [1,*,†], **Shuyao Shang** [1,*], **Jiahe Wang** [1], **Zitong Zhou** [1], **Liang Tan** [1]
Junhan Zeng [1], **Ruizhi Li** [1], **Junyan Li** [1], **Yu Liu** [4], **Xiao Yang** [5], **Yong Li** [5]
Jun Zhu [5], **Hongsheng Li** [2,3], **Tieniu Tan** [1], **Lue Fan** [1,†,✉], **Zhaoxiang Zhang** [1,✉]

[1] CASIA [2] SLAI [3] CUHK [4] AMAP [5] THU

[*] Equal Contribution [†] Project Leaders [✉] Corresponding Authors

Controllable video generation models are increasingly being developed as world models. Accordingly, evaluating them in this role extends beyond the *apparent appearance* of generated videos to the *inherent reactivity* of the worlds they depict: the ability to infer from the scene state how the world should react and to generate plausible consequences not explicitly described in the input. Yet existing benchmarks mainly assess visual quality or explicit instruction fulfillment by checking whether requested actions and interaction outcomes are realized, leaving inherent reactivity underexamined. We introduce **WorldExam**, a hierarchical diagnostic benchmark spanning four levels: Visual Quality, Control Adherence, Spatial Consistency, and World Reactivity. It comprises 1,474 cases across eight dedicated tasks and supports unified evaluation of camera-, action-, and language-driven model paradigms. The World Reactivity level evaluates scene-conditioned reactions and goal-directed behaviors beyond what is explicitly specified in the input. Evaluation of 20 representative models reveals a clear capability split. Camera-driven models excel at camera control, but their interfaces do not support dynamic interaction; action-driven models control subjects more precisely but often leave the world unresponsive; and language-driven models perform better on interaction but follow complex controls less faithfully. No model combines broad task coverage with consistently strong performance, showing that high visual quality and explicit instruction fulfillment do not guarantee inherent reactivity.

Project Website: https://WorldExam.github.io

1 Introduction

Controllable video generation models are increasingly being developed as world models rather than standalone clip generators [3, 5–7, 12, 38, 40, 54, 56]. Such models are expected to predict future visual states from an initial observation and control instructions, including camera trajectories, action sequences, and language prompts. Evaluating them in this role extends beyond the *apparent appearance* of generated videos to the *inherent reactivity* of the worlds they depict [19, 55]. When a subject moves onto stairs, its motion should adapt to the terrain; when it approaches an obstacle, the world should show contact, avoidance, or blockage; when it enters another agent's personal space, that agent should respond plausibly. These effects are scene-conditioned consequences rather than direct depictions of the input. Together, they reveal a model's *inherent reactivity*: its ability to infer from the scene state how the world should react and to generate such consequences plausibly.

Recent benchmarks have advanced world-model evaluation beyond perceptual quality to structured layout control [10], unified action interfaces [11, 51, 52, 57, 58], prompt-specified interaction effects [49, 62], and embodied-AI and autonomous-driving applications [22, 36]. As summarized in Tab. 1, they span camera-, action-, and language-driven model paradigms and increasingly cover camera and subject control, scene

Figure 1 Overview of WorldExam. WorldExam is a hierarchical diagnostic benchmark from *apparent appearance* to *inherent reactivity*. It evaluates 1,474 test cases across camera-, action-, and language-driven interfaces, four diagnostic levels, and eight tasks under a unified evaluation pipeline.

revisiting, and interaction outcomes. Complementary benchmarks probe implicit rules, future-state reasoning, and law-specific physical consistency in specialized settings [24, 26, 43, 48]. Yet most benchmarks still assess explicit instruction fulfillment: a desired layout, camera trajectory, action sequence, or interaction consequence is specified in advance, and the model is evaluated on whether the specified outcome is realized. This evaluation is necessary, but it leaves underexamined a model's ability to infer additional consequences implied by the initial state but not described in the instruction.

We introduce **WorldExam**, a hierarchical diagnostic benchmark designed around this distinction, as summarized in Fig. 1. WorldExam represents each controllable behavior as a composition of atomic control units and adapts these units to each model's native interface: SE(3) camera trajectories for camera-driven models, discrete action sequences for action-driven models, and natural-language prompts for language-driven models. For World Reactivity cases, the model-facing instruction specifies only the explicit control or goal, leaving the expected scene-conditioned reactions unstated. This design distinguishes direct fulfillment of a requested outcome from behavior beyond what the input explicitly specifies.

WorldExam organizes evaluation into the four diagnostic levels: *Visual Quality*, *Control Adherence*, *Spatial Consistency*, and *World Reactivity*. We instantiate this hierarchy with eight evaluation tasks: Camera Control, Subject Control, Scene Revisit, Terrain Interaction, Object Interaction, Social Interaction, Physical Reaction, and Goal Completion. *Visual Quality* is measured with task-agnostic metrics; *Control Adherence* is evaluated through Camera Control and Subject Control; and *Spatial Consistency* is evaluated through Scene Revisit. The *World Reactivity* level covers scene-conditioned reactions and goal-directed behaviors. Within this level, four reaction-oriented tasks use control units as triggers while leaving the induced scene-conditioned reactions unstated. Goal Completion extends the same principle to goal-directed behavior: it specifies a high-level goal while leaving the detailed execution steps unstated. For example, a goal to arrange three bolts by height specifies the target layout, but not which object to move first or how to realize the motion frame by frame.

For model-interface compatibility, WorldExam uses two tracks rather than one global ranking. The static-

Table 1 Comparison with representative world-model benchmarks. The table compares supported model paradigms, viewpoints, task coverage, case counts, and evaluated models. **C**, **A**, and **L** denote camera-, action-, and language-driven model paradigms, respectively. † indicates that the instruction specifies the expected interaction consequence; the corresponding WorldExam tasks leave the evaluated reaction unstated.

Benchmark	Model Paradigm	Viewpoint		WorldExam Evaluation Tasks								#Cases	#Models
		First Person	Third Person	Camera Control	Subject Control	Scene Revisit	Terrain Inter.	Object Inter.	Social Inter.	Physical React.	Goal Compl.		
WorldScore [10]	C/L	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	3,000	20
MIND [57]	A	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	250	2
Omni-WorldBench [49]	C/L	✓	✓	✓	✓	✓	✗	✓†	✗	✓†	✗	1,068	18
WorldMark [52]	C/A/L	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	500	6
iWorld-Bench [11]	C/A/L	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	4,900	14
WBench [58]	C/A/L	✓	✓	✓	✓	✓	✗	✓†	✗	✓†	✗	289	20
WorldOlympiad [62]	A/L	✓	✗	✓	✗	✗	✗	✓†	✗	✓†	✗	1,000	8
WorldRoamBench [51]	A	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	600	10
WorldExam (Ours)	C/A/L	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**1,474**	**20**

scene track controls only the camera and is available to all three paradigms. The dynamic-interaction track requires observable subject–environment interaction and is therefore evaluated only on compatible action- and language-driven models. Separating the tracks avoids treating unsupported capabilities as failures or averaging scores obtained under different scene assumptions and task sets.

Our evaluation of 20 representative models reveals clear trade-offs across the four levels and three paradigms. Camera-driven models excel at camera control, but their interfaces do not support dynamic interaction. Action-driven models control subjects more precisely but often leave the world unresponsive. Language-driven models perform better on interaction tasks but follow complex controls less faithfully. These capability splits are obscured by aggregate scores, motivating separate reporting at both level and task granularity.

Our contributions are summarized as follows.

- We extend world model evaluation beyond apparent appearance to *inherent reactivity*: inferring from the scene state how the world should react and generating plausible consequences absent from the input.
- We propose WorldExam, a benchmark of 1,474 cases across eight tasks that supports unified evaluation of camera-, action-, and language-driven model paradigms.
- We evaluate 20 representative models, revealing paradigm-dependent capability splits. No model combines broad task coverage with consistently strong performance, showing that high visual quality and explicit instruction fulfillment do not guarantee inherent reactivity.
- We will publicly release the benchmark data and evaluation toolkit to facilitate systematic evaluation and foster continued progress in the video world model community.

2 Related Work

2.1 Video World Models

Recent video world models increasingly support controllable video generation for gaming, robotics, embodied AI, and open-world simulation. Based on their primary control interfaces, they can be broadly grouped into camera-, action-, and language-driven paradigms. Camera-driven models [3, 9, 13, 39, 56, 59] condition generation on camera trajectories, represented in two main ways. Some approaches, such as ReCamMaster [3] and FantasyWorld [9], inject camera trajectories through learned camera encoders or embeddings, whereas others [13, 39, 56, 59] reconstruct 3D priors from the input and reproject them to target viewpoints; representative methods include NeoVerse [56] and InSpatio-World [39]. Action-driven models [20, 30, 38, 40, 46, 50, 65] generate future frames conditioned on discrete action sequences through keyboard-like interfaces. Among them, WorldPlay [38] and LingBot-World [40] focus on real-time interaction and consistent generation

under direct action control. Language-driven models [1, 5, 12, 17, 32, 35, 44] generate videos from text or image-text prompts, demonstrating advances in semantically complex video generation. Across paradigms, video world models are evolving from short open-loop synthesis toward controllable, persistent, and interactive environment simulation. Heterogeneous interfaces complicate direct comparison, while controllability, long-term memory, and *inherent reactivity* remain key challenges.

2.2 Video World Model Benchmarks

A growing body of benchmarks evaluates complementary aspects of video world modeling. Some emphasize perceptual and temporal quality [14, 15, 27, 28, 63]; others target compositionality, world knowledge, implicit rules, and future-state reasoning [8, 26, 37, 48]. Physics-oriented benchmarks [4, 18, 24, 31, 43, 47, 53] diagnose law-specific dynamics, geometric consistency, and generalization under physical interactions; embodied benchmarks [16, 19, 21, 29, 33, 36, 55, 60] evaluate action fidelity, physical executability, planning utility, reliability, and trustworthiness; and autonomous-driving benchmarks [2, 22, 64] emphasize ego-action control, trajectory plausibility, safety, and downstream driving utility. Beyond these settings, general benchmarks [10, 11, 49, 51, 52, 57, 58, 61, 62] evaluate interactive world models across varied scenes and interfaces.

Among these general benchmarks, WorldScore [10] evaluates camera-trajectory-based layout control and geometric consistency, while MIND [57] focuses on action control and closed-loop revisit consistency. WorldMark [52] and iWorld-Bench [11] improve cross-model comparison through standardized or unified action representations. Omni-WorldBench [49] evaluates prompt-specified interaction outcomes, affected and unaffected entities, and intermediate causal state transitions; WBench [58] extends evaluation to multi-turn navigation, subject actions, event editing, and perspective switching; and WorldOlympiad [62] probes long-horizon interaction and physics. WorldRoamBench [51] further couples long-horizon action-conditioned generation with diagnostics of controllability, visual drift, mechanics, optics, 3D consistency, and memory. Collectively, these benchmarks substantially broaden interactive evaluation, but most still center on explicit instruction fulfillment by checking whether specified controls or interaction outcomes are realized. In contrast, WorldExam adapts atomic control units to each model's native interface and evaluates *inherent reactivity* through scene-conditioned reactions and goal-directed behaviors beyond what the input explicitly specifies.

3 WorldExam

WorldExam supports unified evaluation of video world models with different control interfaces. We formulate a video world model as a function $f : \mathcal{I} \times \mathcal{C} \to \mathcal{V}$, where $\mathcal{I}$ is the initial image, $\mathcal{C}$ is the model-facing input instruction and $\mathcal{V}$ is the generated video. We consider three common paradigms: *camera-driven* models take camera trajectories in SE(3), *action-driven* models take discrete action sequences over {W (move forward), S (move backward), A (move left), D (move right), ↑ (tilt up), ↓ (tilt down), ← (pan left), → (pan right), ∅ (stop)}, and *language-driven* models take natural-language prompts.

To compare these paradigms, WorldExam uses interface adaptation to map a shared case to each model's native interface. WorldExam represents controllable behavior as an ordered composition of atomic control units, such as moving forward (W) and then panning right (→), and adapts this control intent into an SE(3) camera trajectory, a discrete action sequence, or a natural-language prompt. Under this setup, WorldExam first defines a four-level diagnostic hierarchy (Sec. 3.1) and then instantiates it through eight evaluation tasks (Sec. 3.2). We next describe the test case curation pipeline (Sec. 3.3) and the benchmark statistics (Sec. 3.4). Metric definitions and scoring protocols are described in Sec. 4.

3.1 Toward World Reactivity: A Four-Level Diagnostic Hierarchy

WorldExam organizes world-model evaluation into four diagnostic levels of increasing scope: *Visual Quality*, *Control Adherence*, *Spatial Consistency*, and *World Reactivity*. *Visual Quality* measures the video's *apparent appearance*, including perceptual plausibility, temporal stability, and aesthetic quality. *Control Adherence* measures whether the controlled camera or subject follows the input control. *Spatial Consistency* measures whether the model preserves a coherent world when the camera revisits a previously observed viewpoint.

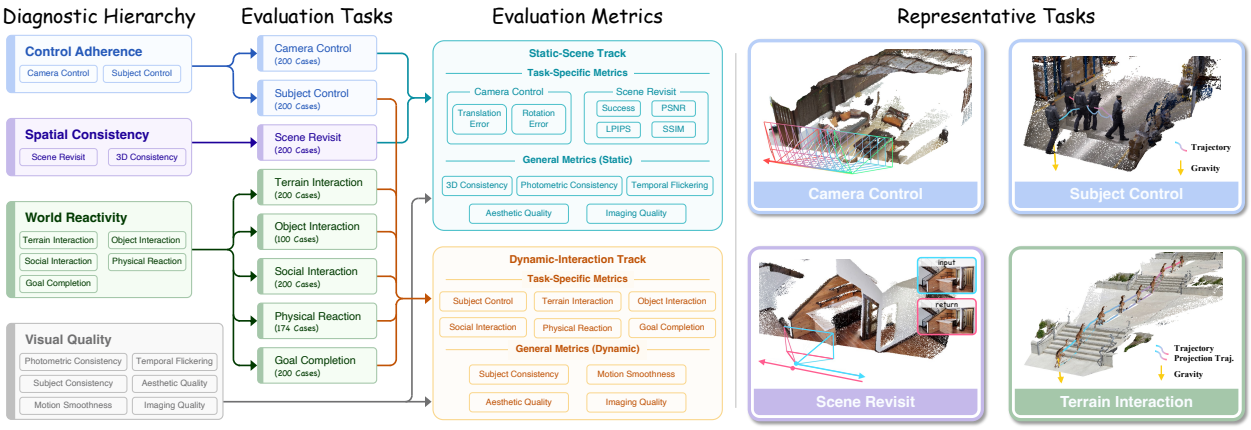


Figure 2 WorldExam taxonomy, tracks, and metrics. Four diagnostic levels map to eight evaluation tasks, which are assigned to static-scene or dynamic-interaction tracks according to scene assumptions and model applicability. Each track reports task-specific and general metrics. Representative examples illustrate the geometry-based evaluations.

By contrast, the *World Reactivity* level evaluates scene-conditioned reactions and goal-directed behaviors beyond what is explicitly specified in the input. In reaction-oriented cases, a control specifies the initiating behavior but leaves its scene-conditioned consequences unstated, which the model must infer from the scene. For example, a move-forward control specifies the subject's direction but not how its motion should adapt to the terrain. If there is an obstacle or a nearby agent in its motion path, the subject may stop or avoid it, another agent may yield, or an object may move on contact. In goal-directed cases, a high-level goal specifies the desired target, and the model needs to infer from the initial scene how to realize it frame by frame.

Although the four levels form a diagnostic progression, strong *Visual Quality*, *Control Adherence*, and *Spatial Consistency* do not guarantee successful scene-conditioned reactions or goal execution. Conversely, success on *World Reactivity* does not compensate for visual artifacts, control errors, or spatial drift. WorldExam therefore reports the four levels separately to localize failures in generation quality, explicit control, spatial persistence, and behavior that must be inferred from the scene.

3.2 From Diagnostic Levels to Evaluation Tasks

Fig. 2 shows how the four diagnostic levels are instantiated. *Visual Quality* uses task-agnostic metrics across all videos, while the other three levels are instantiated by eight tasks. *Control Adherence* includes **Camera Control** and **Subject Control**, while *Spatial Consistency* uses **Scene Revisit**. *World Reactivity* includes four reaction-oriented tasks, **Terrain Interaction**, **Object Interaction**, **Social Interaction**, and **Physical Reaction**, plus **Goal Completion** for goal-directed execution.

The tasks are reported through two tracks rather than a single global score to avoid penalizing models for tasks their interfaces do not support. The *static-scene track* contains Camera Control and Scene Revisit and is available to all three model paradigms because each interface can express camera motion. The *dynamic-interaction track* contains Subject Control and the five *World Reactivity* tasks and applies only to compatible action- and language-driven models; Goal Completion is language-only.

Control Adherence. **Camera Control** tests whether the generated camera motion follows the prescribed controls. Each case composes one to three atomic camera controls from {W, S, A, D, ↑, ↓, ←, →}, assigns each an execution-time fraction, and executes them in order over the assigned intervals. **Subject Control** applies the same construction to a designated third-person subject using {W, S, A, D}.

Spatial Consistency. **Scene Revisit** tests the model's spatial memory of the initial observation. Each case uses a round-trip camera trajectory formed by an outgoing control and its inverse, such as "move left" followed by "move right", or "tilt up" followed by "tilt down". After moving away, the camera should return to the initial viewpoint while the returned view preserves the scene's geometry, appearance, and content.

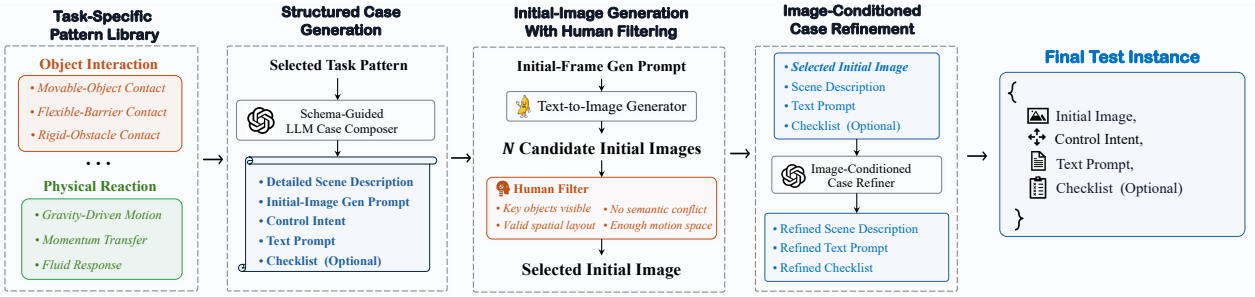


Figure 3 Test case curation pipeline. For the six dynamic-interaction tasks, a task pattern is expanded into a structured draft, candidate initial images are generated and human-filtered, and the scene description, text prompt, and optional checklist are refined against the selected image before the case is finalized.

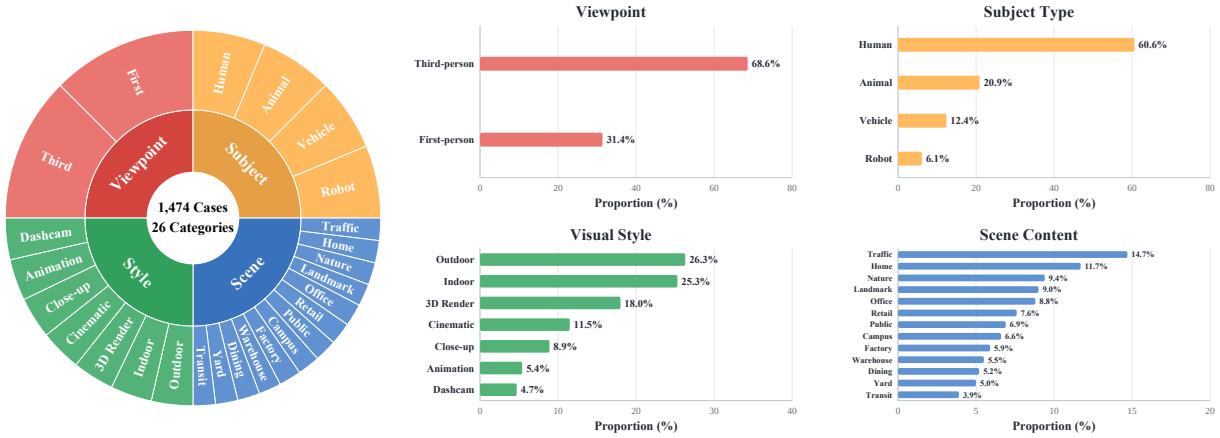
World Reactivity. The four reaction-oriented tasks pair an initial scene with a single atomic subject control. **Terrain Interaction** places stairs, slopes, bridges, trenches, or other structured terrain along the controlled subject's path. The input specifies only the horizontal motion direction, while the model must infer how the subject should adapt its height and maintain contact with the terrain. **Object Interaction** places a movable, flexible, or rigid target along the subject's path so that the subject, one of its body parts, or a carried tool is expected to make contact with it. It evaluates whether the target produces an immediate type-appropriate response, such as motion when loose, deformation when flexible, or blockage when rigid, without interpenetration. **Social Interaction** places other agents along the subject's path or within its social distance, creating an imminent local conflict. It evaluates whether the affected agents respond plausibly through avoidance, yielding, stopping, or changing path. **Physical Reaction** tests whether a dynamic process unfolds over time according to physical regularities, including gravity, friction, momentum transfer, constrained motion, fluid response, and pendulum-like swinging. Each case uses one control from {W, S, A, D, ∅ (stop)}. A motion control may trigger the process, whereas ∅ is used when the initial scene is expected to evolve autonomously without subject motion. Although Object Interaction and Physical Reaction may both involve contact, the former targets the immediate type-conditioned response of a designated object, whereas the latter targets the temporal evolution of a physical process. **Goal Completion** is language-only and provides a high-level goal together with an initial scene containing relevant entities, distractors, preconditions, and constraints. Unlike the four reaction-oriented tasks, it uses no atomic control sequence or execution-time fractions. The input may state necessary subgoals or ordering constraints. The model should ground the goal in the initial scene, select the correct entities, ignore distractors, and produce coherent execution steps toward the desired target frame by frame. Sec. C provides examples of all eight tasks and representative checklists.

3.3 Test Case Curation Pipeline

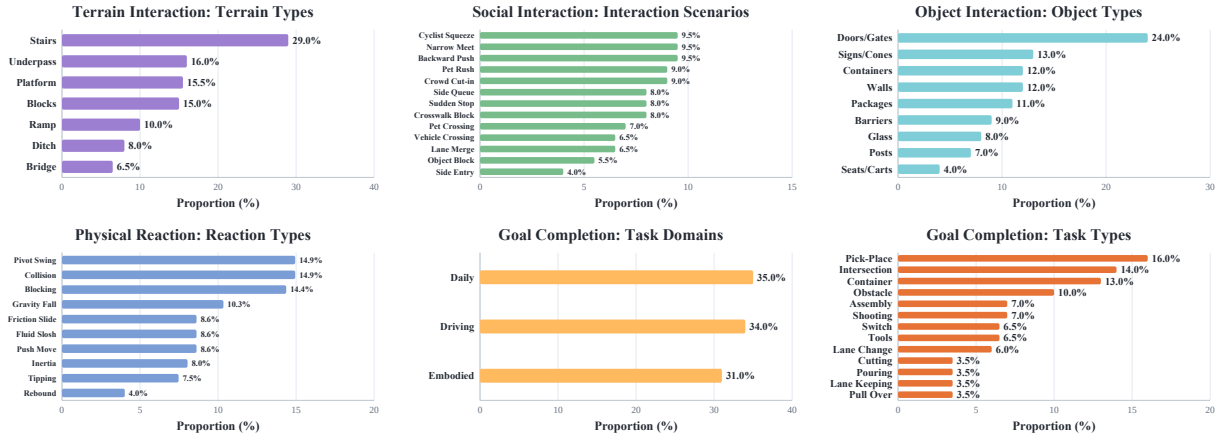
WorldExam constructs cases differently for the two tracks. For static-scene Camera Control and Scene Revisit, we pair suitable first-person scenes from existing datasets [25, 41, 57] with compositions of atomic control units. For dynamic-interaction Subject Control and the five World Reactivity tasks, the pipeline in Fig. 3 constructs initial scenes supporting the intended behavior and evaluation.

For each dynamic-interaction task, a task-specific pattern library defines the intended semantic coverage over subject motion, structured terrain, object contact, social conflict, physical processes, or goal-directed situations. After a pattern is sampled, a schema-guided LLM case composer expands it into a structured draft containing a detailed scene description, an initial-image generation prompt, a control intent or high-level goal, and a draft text prompt for language-driven models. For Object Interaction, Social Interaction, Physical Reaction, and Goal Completion, the draft also includes a case-specific checklist of observable evaluation criteria. In each World Reactivity case, the model-facing input specifies only the explicit control or high-level goal; the scene-conditioned reaction or detailed execution process remains unstated.

The initial-image generation prompt is used only to synthesize N candidate initial images. Human filtering retains candidates in which the relevant entities are visible, the spatial layout supports the intended behavior or event, the image is consistent with the draft, and sufficient motion space remains for the continuation.



(a) Dataset composition.



(b) World Reactivity composition.

Figure 4 Benchmark statistics. The top panel summarizes distributions by viewpoint, subject type, visual style, and scene content. The bottom panel shows terrain types, social-interaction scenarios, object types, physical-reaction types, and Goal Completion domains and task types.

Candidates that already depict the evaluated event or desired target, hide relevant entities, or make the intended behavior physically infeasible are discarded. The selected initial image I^0 therefore provides a concrete pre-event state from which the intended behavior, reaction, or goal-directed execution can unfold.

An image-conditioned case refiner then revises the draft to match I^0 while preserving the sampled pattern and intended control or goal. It updates entity references, spatial relations, the scene description, the text prompt, and the optional checklist so that all referenced entities and preconditions are grounded in the selected image. For tasks evaluated using checklists, the final checklist $L = \{\ell_k\}_{k=1}^{K}$ is fixed at this stage. The finalized case consists of I^0, the control intent or high-level goal, the grounded text prompt, the optional checklist.

3.4 Benchmark Statistics

As shown in Fig. 2, WorldExam contains 1,474 cases across eight evaluation tasks. Fig. 4 summarizes both the overall dataset composition and the task-specific composition of the five *World Reactivity* tasks.

At the dataset level, the cases span first- and third-person viewpoints with first-person viewpoints accounting for 31.4% of the benchmark and providing substantial egocentric coverage. The subject taxonomy spans humans, animals, vehicles, and robots, while the visual-style taxonomy mixes outdoor and indoor real scenes with 3D renderings, cinematic footage, close-up views, animation, and dashcam videos. Scene content is also deliberately broad: no single scene type dominates the benchmark, and the largest category, traffic scenes,

accounts for only 14.7% of the cases.

Within the five *World Reactivity* tasks, the cases are further distributed across task-specific semantic subcategories, including terrain types, social-interaction scenarios, object types, physical-reaction types, and Goal Completion domains and task types. No single subcategory accounts for more than 35% of its corresponding task. This coverage reduces dependence on any one visual or semantic template and supports task-specific analysis across diverse scene-conditioned reactions and goal-directed situations. Representative cases across these dimensions are shown in Sec. A.

4 Evaluation Protocol and Metrics

For Camera Control, Scene Revisit, Subject Control, and Terrain Interaction, we lift each generated video into 3D with a geometry reconstruction model and evaluate it in the reconstructed space. Camera Control and Scene Revisit use the recovered camera trajectories, whereas Subject Control and Terrain Interaction use the recovered 3D subject trajectories and terrain geometry. For the remaining four tasks, we use GPT-5.5 as the vision-language model (VLM) judge to score generated videos against predefined case-specific checklists. Each track reports task-specific metrics together with task-agnostic general metrics for visual quality.

4.1 Static-Scene Track

Given generated frames $V = \{I_t\}_{t=1}^{T}$, we use VGGT-Ω [45] to estimate camera poses, intrinsics, and depths.

Camera Control. Each case specifies an ordered sequence $\{(a_i, \rho_i)\}_{i=1}^{N} (1 \le N \le 3)$. Here, $a_i \in$ {W, S, A, D, ↑, ↓, ←, →} is an atomic control unit, and ρ_i is its execution-time fraction, with $\sum_i \rho_i = 1$. For camera- and action-driven interfaces, we allocate $n_i = [\rho_i T]$ frames to the i-th control using nearest-integer rounding, and adjust the allocation to ensure $\sum_i n_i = T$. From each generated video, we recover a frame-wise camera trajectory $\{(\mathbf{R}_t, \mathbf{t}_t)\}_{t=1}^{T}$. Because the three interfaces specify camera motion differently, we construct the model-facing input and evaluation reference separately for each interface.

For camera-driven models, controls are composed sequentially starting from the initial camera pose, with each subsequent control applied relative to the endpoint pose of the previous control. These endpoints serve as keyframes, which we interpolate over the allocated frame intervals to obtain a frame-wise input trajectory in SE(3). This input trajectory also serves directly as the frame-wise reference trajectory $\{(\mathbf{R}_t^*, \mathbf{t}_t^*)\}_{t=1}^{T}$. Before comparison, we express both the recovered and reference trajectories relative to their respective first-frame poses. We then compute the translation and rotation errors (e_t, e_r) between the two trajectories as

$$e_t = \min_{s \ge 0} \frac{1}{T} \sum_{t=1}^{T} \| s\mathbf{t}_t - \mathbf{t}_t^* \|_2 , \qquad e_r = \frac{180}{\pi} \frac{1}{T} \sum_{t=1}^{T} \arccos\left(\frac{\mathrm{tr}\big(\mathbf{R}_t (\mathbf{R}_t^*)^\top\big) - 1}{2} \right). \tag{1}$$

The nonnegative scale s resolves the translation-scale ambiguity of monocular camera reconstruction, making e_t scale-invariant, while $180/\pi$ converts e_r from radians to degrees. In implementation, the argument of arccos is clipped to $[-1, 1]$ for numerical stability.

Action-driven models map each control to the model's native discrete action and assign the corresponding n_i frames to the i-th action. Language-driven models instead verbalize each control, join the resulting motion phrases in order with "then," and prepend the instruction to the scene description; for example, "W" followed by "→" becomes "The camera moves forward, then pans right. [Scene description]." This prompt preserves the control order but does not specify the duration of each control.

Unlike camera-driven interfaces, action- and language-driven interfaces do not specify an exact camera trajectory in SE(3), so we evaluate their recovered trajectories against control-level references segment by segment. For action-driven models, the frame ranges assigned to the discrete actions directly define the segment boundaries. For language-driven models, we instead partition the recovered trajectory into N segments by applying dynamic-programming-based change-point detection [42] to frame-to-frame changes in translation and rotation. The resulting segments are matched in temporal order to the N atomic controls.

Within each segment, we express the recovered camera poses relative to the first frame, so that the segment starts from the identity pose. The assigned control determines whether the reference motion is a translation

or a rotation. For a translation control, we linearly interpolate the reference translation from $\mathbf{0}$ to a unit vector $\mathbf{u}_i$ in the prescribed direction, while keeping $\mathbf{R}^* = \mathbf{I}$ throughout. The unit displacement is sufficient because e_t is invariant to translation scale. For a rotation control, we set $\mathbf{t}^* = \mathbf{0}$ and construct $\mathbf{R}^*$ using the prescribed axis and direction; its translation error is computed as the mean per-frame $\|\mathbf{t}_t\|_2$ without scale alignment. Because the interface does not specify a rotation angle, the reference angle is linearly interpolated from zero to the total angle recovered within the segment. The segment-level references therefore evaluate the prescribed direction and motion progression without imposing a fixed magnitude.

For each segment i, we compute $(e_{t,i}, e_{r,i})$ using the error definitions above. A translation segment is assigned the maximum translation error $e_{t,i} = 0.5$ if its displacement is below 5% of the largest segment displacement in the same video or if its net motion is not aligned with the prescribed direction. A rotation segment is assigned the maximum rotation error $e_{r,i} = 15°$ if its total rotation angle is below 5° or is opposite to the prescribed direction. Finally, we obtain the video-level errors (e_t, e_r) by averaging the corresponding segment-level errors using the number of frames in each segment as weights.

Across all interfaces, we normalize the resulting errors as $s_t = \max(0, 1 - e_t/0.5)$ and $s_r = \max(0, 1 - e_r/15)$, where e_r is measured in degrees, and report their geometric mean, $S_{\text{cam}} = 100\sqrt{s_t s_r}$. Thus, a high Camera Control score requires the recovered camera motion to follow the prescribed directions and temporal progression.

Scene Revisit. Scene Revisit evaluates two requirements after a round-trip camera motion: returning the camera to its initial pose and preserving the initial scene in the returned view. Each case pairs an outgoing control a_1 with its inverse $a_2 = a_1^{-1}$. We set their execution-time fractions to $\rho_1 = 0.4$ and $\rho_2 = 0.6$, reserving a longer temporal window for the return motion so that the model has sufficient opportunity to reach the initial viewpoint. We adapt this pair to the three interfaces as in Camera Control.

Let $P_t = (\mathbf{R}_t, \mathbf{t}_t)$ denote the recovered camera pose. For camera- and action-driven models, $\mathcal{T}_{\text{return}}$ is the frame range allocated to a_2; for language-driven models, whose prompt does not specify control duration, it begins at 40% of the video. Because the camera may return before the video ends, we search this entire segment and select the frame whose recovered pose is closest to the initial pose:

$$t_{\text{rev}} = \underset{t \in \mathcal{T}_{\text{return}}}{\arg\min}\ d(P_t, P_1). \tag{2}$$

For translation round trips, $d(P_t, P_1) = \|\mathbf{t}_t - \mathbf{t}_1\|_2$; for rotation round trips, d is the relative rotation angle between $\mathbf{R}_t$ and $\mathbf{R}_1$. A translation revisit succeeds when this minimum distance is within 10% of the maximum displacement reached during the outgoing segment; a rotation revisit succeeds when its minimum angular distance is below 5°. Averaging this binary result over all cases gives Revisit Success $S_{\text{succ}} \in [0, 1]$.

We then compare the input image with the selected revisit frame using PSNR, LPIPS, and SSIM. Selecting the frame by recovered pose rather than using the final frame makes this appearance comparison insensitive to small differences in return timing. After averaging over cases, we normalize the three appearance metrics as $s_{\text{P}} = \min(\text{PSNR}/25, 1)$, $s_{\text{L}} = 1 - \text{LPIPS}$, and $s_{\text{S}} = \text{SSIM}$, and combine them with Revisit Success:

$$S_{\text{rev}} = 100\sqrt{S_{\text{succ}} \cdot \frac{s_{\text{P}} + s_{\text{L}} + s_{\text{S}}}{3}}. \tag{3}$$

The geometric aggregation gives a high Scene Revisit score only when the camera both returns to the initial viewpoint and recovers a consistent view.

General metrics. Across all static-scene videos, we additionally report five task-agnostic general metrics. 3D Consistency adapts the metric of WorldScore [10] to VGGT-Ω outputs. Using the recovered geometry, we project valid pixels from a source frame to a nearby target and back, then measure the cycle reprojection error. Photometric Consistency measures the forward-backward optical-flow cycle error between neighboring frames using average endpoint error (AEPE). Temporal Flickering, Aesthetic Quality, and Imaging Quality are adapted from VBench [14]. These metrics summarize whether the static-scene generations are geometrically stable, temporally coherent, and visually plausible.

4.2 Image-Space Displacement Alignment for Camera-Driven Models

Identical translation values in an input SE(3) trajectory can induce different image-space displacements across camera-driven models because their pose-conditioning interfaces interpret the translation magnitude

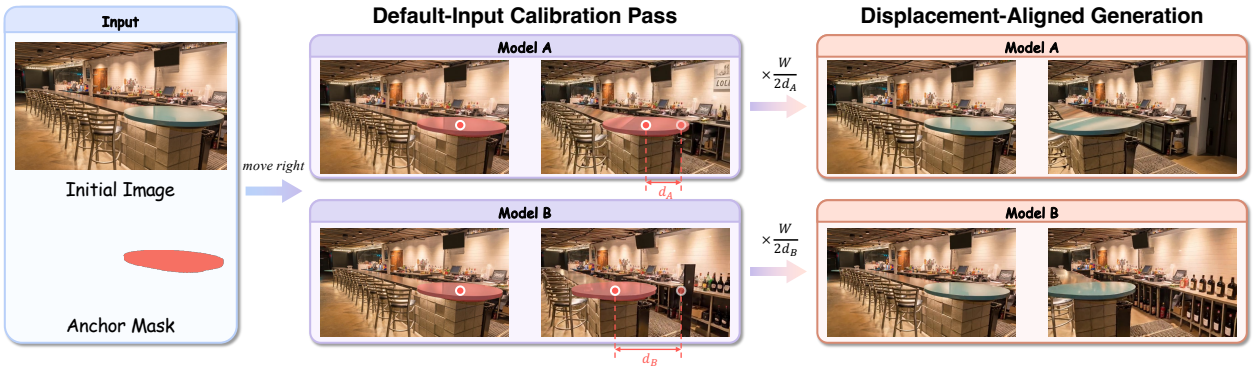


Figure 5 Image-space displacement alignment for camera-driven models. Given an initial image and anchor mask for case c, WorldExam measures model m's image-space displacement $d_{m,c}$ in a default-input calibration pass and scales its input translations by $k_{m,c} = W/(2d_{m,c})$ to align displacement across camera-driven models.

differently. Larger displacements expose more novel-view content and increase the difficulty of generating controlled, spatially consistent videos. By increasing the input translation multiplier to induce progressively larger image-space displacements, the results in Tab. 3 and Fig. 7 show corresponding declines in Camera Control, Scene Revisit, and general-metric performance.

The scale-invariant translation error in Sec. 4.1 does not address this difference. Its scalar s is fitted after generation only to resolve the coordinate-scale mismatch between the recovered and reference trajectories; it neither changes the input trajectory nor normalizes the image-space displacement in the generated video. We therefore calibrate input translations before generation to align image-space displacement (Fig. 5).

For each model m and case c, we estimate a translation calibration factor using the anchor mask provided on the initial frame. We first generate a calibration video with the model's default input translation magnitude, apply a horizontal camera control, either "move left" or "move right", and track the anchor through the video using SAM2 [34]. Let $d_{m,c}$ denote the horizontal image-space displacement of the tracked mask center, and let W denote the frame width. We set the target displacement to $d^* = W/2$ and compute $k_{m,c} = d^*/d_{m,c} = W/(2d_{m,c})$. For the final generation used in evaluation, we multiply the translation components of the default input trajectory by $k_{m,c}$ while leaving its rotations unchanged.

4.3 Dynamic-Interaction Track

Subject Control. Subject Control uses the same ordered-control construction as Camera Control, with $a_i \in \{\text{W, S, A, D}\}$ applied to a designated subject. Action-driven models receive native discrete subject actions over the allocated frame ranges. For language-driven models, we verbalize the ordered controls together with the designated subject and prepend the resulting instruction to the scene description; for example, "W" followed by "A" becomes "The [subject] moves forward, then the [subject] moves left. [Scene description]."

As illustrated in Fig. 6, SAM2 [34] tracks the designated subject from a first-frame mask, and VGGT-Ω [45] lifts the tracked pixels into 3D to recover the subject trajectory. The horizontal-region mask estimates gravity and the horizontal plane; projecting the initial camera's viewing direction onto this plane defines the forward reference direction, with the other directions derived analogously.

We evaluate the recovered subject trajectory segment by segment. The action-driven segment boundaries follow the frame ranges assigned to the discrete subject actions, whereas the N language-driven segments are inferred by applying the same change-point procedure as in Camera Control to frame-to-frame subject displacement. Within each segment, we translate the recovered trajectory so that its first-frame subject position is the origin. The associated atomic control unit selects a horizontal unit direction $\mathbf{u}_i$, and the reference trajectory is linearly interpolated from the origin to $\mathbf{u}_i$. We fit a post-generation translation scale s between the recovered and reference trajectories, as in Camera Control, and compute the segment-level translation error $e_{t,i}$. A segment whose net displacement is below 0.5% of the reconstructed scene scale or whose motion is misaligned with the prescribed direction receives the maximum error $e_{t,i} = 0.5$. Finally, the error e_t is obtained by

Figure 6 Geometry-based evaluation of Subject Control and Terrain Interaction. (a) SAM2 and VGGT-Ω recover the subject trajectory, while the horizontal-region mask estimates gravity and the horizontal plane used to define the control directions. (b) A subject-free image restores the complete terrain geometry, onto which the trajectory is projected along gravity to obtain corresponding terrain trajectory.

frame-count-weighted averaging, and the Subject Control score is $S_{\text{sub}} = 100 \max(0, 1 - e_t/0.5)$.

Terrain Interaction. Unlike Subject Control, Terrain Interaction uses a single atomic control unit to induce a subject-terrain interaction. For evaluation, we additionally provide a subject-free terrain image to recover the complete terrain geometry. We then project the recovered 3D subject trajectory along gravity onto this geometry to obtain the corresponding 3D terrain trajectory.

During evaluation, we first compute the Subject Control score from the horizontal component of the subject trajectory and use it as a gating check; cases that fail this check are considered not to follow the control and receive a Terrain Interaction score of zero. For cases that pass this check, we extract local extrema and the endpoint from the height of the terrain trajectory as evaluation points. The Terrain Interaction score is the ratio of the number of evaluation points at which the subject and terrain trajectories exhibit consistent height changes to the total number of evaluation points.

Checklist-Based World Reactivity Evaluation. Considering that the remaining four tasks require semantic and causal judgments that cannot be captured by recovered trajectories, we evaluate them with a VLM judge against the case-specific checklist $L = \{\ell_k\}_{k=1}^{K}$ constructed in Sec. 3.3. Each checklist covers the initiating condition, the resulting reaction or goal execution progress, and invalid outcomes. At evaluation time, the VLM receives 10 temporally ordered frames uniformly sampled from the generated video, together with the checklist. Using a task-specific prompt, the judge returns one binary decision per item; contradicted, missing, ambiguous, off-screen, or otherwise unverifiable evidence is counted as unsatisfied. The case score is

$$S_{\text{check}}(V, L) = \frac{100}{K} \sum_{k=1}^{K} \mathbb{I}[\ell_k \text{ is satisfied in } V]. \tag{4}$$

Object Interaction. The checklist verifies that contact occurs with the designated object, precedes and causes the reaction, and produces a type-consistent outcome without interpenetration. It also checks that the direction and extent of the reaction remain consistent with the contact.

Social Interaction. The checklist verifies that the controlled motion creates the intended conflict and that at least one visible affected agent makes a timely adjustment attributable to the controlled subject. Unchanged, delayed, unrelated, or physically implausible responses are counted as failures.

Table 2 Static-scene track evaluation. Task averages the Camera Control and Scene Revisit scores, **General** averages the five general metrics, and **Overall** averages Task and General scores. Down ↓ and up ↑ arrows indicate that lower and higher values are better, respectively. The best and second-best results per paradigm are bold and underlined.

Model	Camera Control			Scene Revisit					3D Cons.	Photo. Cons.	Temp. Flick.	Aesth. Quality	Imag. Quality	Average		
	T. Err.↓	R. Err.↓	Score↑	Success↑	PSNR↑	LPIPS↓	SSIM↑	Score↑						Task	General	Overall
Camera-driven																
TrajectoryCrafter [59]	0.14	1.01	80.32	**1.000**	19.08	0.240	0.545	83.03	96.98	62.83	92.56	52.17	67.03	81.68	74.31	78.00
ReCamMaster [3]	0.47	3.91	38.64	0.815	16.02	0.353	0.415	68.01	**99.33**	**81.97**	95.52	54.58	73.45	53.33	**80.97**	67.15
Voyager [13]	0.31	4.29	56.19	0.995	16.88	0.388	0.466	76.27	89.54	27.43	93.04	53.59	63.69	66.23	65.46	65.84
FantasyWorld [9]	1.04	7.82	18.46	0.560	15.41	0.323	0.374	55.79	98.57	75.60	**95.91**	**57.11**	**73.98**	37.12	80.23	58.68
NeoVerse [56]	**0.01**	**0.60**	**97.33**	**1.000**	**21.72**	**0.141**	**0.662**	**89.25**	98.19	70.83	93.53	52.72	72.19	**93.29**	77.49	**85.39**
InSpatio-World (1.3B) [39]	0.09	1.24	85.94	**1.000**	20.79	0.224	0.606	85.90	98.44	64.88	93.68	53.78	73.60	85.92	76.88	81.40
Action-driven																
Hunyuan-GameCraft [20]	0.09	11.95	41.33	0.385	13.36	0.542	0.367	41.77	93.61	53.93	93.91	54.97	71.69	41.55	73.62	57.59
Astra [65]	0.31	7.92	32.59	0.365	12.20	0.601	0.309	38.15	96.46	78.41	96.38	51.71	71.78	35.37	78.95	57.16
WorldPlay [38]	**0.04**	**1.04**	**92.74**	0.790	**18.42**	**0.271**	**0.531**	**72.51**	98.69	81.06	95.90	53.92	73.31	**82.63**	80.58	**81.61**
Yume 1.5 [30]	0.12	3.23	75.67	0.255	12.37	0.615	0.359	32.45	98.44	67.53	95.18	53.42	75.28	54.06	77.97	66.02
LingBot-World [40]	0.11	6.22	58.19	0.605	15.58	0.341	0.406	58.35	99.59	81.52	**96.53**	**59.87**	75.00	58.27	82.50	70.39
Infinite-World [50]	0.17	2.35	71.70	0.645	14.31	0.388	0.345	57.34	**99.91**	**92.09**	95.99	56.27	**76.41**	64.52	**84.13**	74.33
Matrix-Game 3.0 [46]	0.06	4.36	76.36	**0.860**	13.82	0.485	0.372	64.25	95.59	30.14	93.40	48.80	73.39	70.30	68.26	69.28
Language-driven																
Kling 2.5 [17]	0.19	6.90	50.18	0.325	**13.54**	0.532	0.381	38.81	**99.87**	**87.87**	**97.56**	56.28	75.70	44.50	**83.46**	63.98
Veo 3.1 [12]	0.27	8.22	40.83	0.447	12.30	0.594	0.345	43.05	99.06	73.61	95.43	55.49	76.57	41.94	80.03	60.99
Hailuo 2.3 [32]	**0.15**	**5.20**	**63.29**	**0.505**	13.52	**0.519**	**0.387**	**48.70**	99.38	79.88	95.33	**56.30**	75.52	**55.99**	81.28	**68.64**
Wan 2.6 I2V [44]	0.17	6.12	57.72	0.495	13.28	0.527	0.353	47.32	99.49	69.42	94.13	53.93	**77.53**	52.52	78.90	65.71
Seedance 1.5 [35]	0.20	7.69	49.18	0.375	12.28	0.602	0.342	39.23	97.31	56.60	94.81	54.45	74.88	44.21	75.61	59.91
Vidu Q3 [5]	0.24	8.47	42.75	0.360	12.65	0.586	0.351	39.05	99.43	71.28	95.01	55.38	76.73	40.90	79.57	60.24
HappyHorse 1.0 [1]	0.18	5.62	58.29	0.420	12.48	0.564	0.352	42.45	99.62	74.17	95.02	55.56	77.00	50.37	80.27	65.32

Physical Reaction. The checklist verifies the timing and cause of the process, its evolution under the relevant physical regularity, and the preservation of required supports, attachments, contacts, and constraints. Freezing, premature onset, interpenetration, broken attachments, or unexplained energy are counted as failures.

Goal Completion. The checklist separately evaluates correct grounding, intermediate progress, compliance with stated ordering and scene-dependent constraints, and final completion. Scores credit partial progress and accept alternative executions that reach the desired target under the same observable requirements.

General metrics. The dynamic-interaction track separately reports four VBench metrics [14]: Subject Consistency and Motion Smoothness for feature and temporal consistency, and Aesthetic Quality and Imaging Quality for visual appeal and frame-level quality. We omit 3D Consistency, Photometric Consistency, and Temporal Flickering because valid motion and state changes disrupt their geometric and optical-flow correspondences.

5 Experiments

5.1 Experimental Setup

We evaluate 20 representative video world models: 6 camera-driven, 7 action-driven, and 7 language-driven models. For each case, we construct the model-facing input in the model's native format using the interface adaptation described in Sec. 3 and evaluate the generated video with the protocols in Sec. 4. All 20 models are evaluated on the *static-scene track*. The *dynamic-interaction track* requires observable third-person subject-scene interaction and therefore includes WorldPlay [38], LingBot-World [40], and all seven language-driven models. The other five action-driven models either do not support third-person subject control or cannot control a visible third-person subject reliably, and are therefore excluded from the dynamic-interaction track. Camera-driven models are excluded because their interfaces control only the camera, and Goal Completion is evaluated only on language-driven models. To avoid compromising model performance, we use each model's default resolution, video length, and other inference settings whenever applicable. For streaming models with flexible generation lengths, we constrain the output to 100–200 frames to prevent quality drift in excessively long generations from biasing the evaluation. Some closed-source commercial language-driven systems apply proprietary prompt enhancement before video generation. Consistent with the functional formulation in Sec. 3, we retain such default preprocessing as part of the native end-to-end pipeline. Detailed inference settings and task eligibility are provided in Sec. B; unsupported tasks are marked with dashes in the result tables.

Table 3 Effect of the input translation multiplier on NeoVerse. The multiplier is applied only to the translation components of the input SE(3) trajectory; rotations remain unchanged. Metrics and the Task, General, and Overall aggregates follow Tab. 2.

Translation Multiplier	Camera Control			Scene Revisit					3D Cons.	Photo. Cons.	Temp. Flick.	Aesth. Quality	Imag. Quality	Average		
	T. Err.↓	R. Err.↓	Score↑	Success↑	PSNR↑	LPIPS↓	SSIM↑	Score↑						Task	General	Overall
0.10×	0.002	0.43	98.25	1.000	22.68	0.128	0.704	90.98	99.91	80.17	95.41	54.06	72.53	94.62	80.42	87.52
0.50×	0.003	0.51	97.84	1.000	22.17	0.134	0.683	90.11	99.54	76.33	94.21	53.39	72.47	93.98	79.19	86.59
0.75×	0.004	0.56	97.57	1.000	22.00	0.137	0.676	89.80	99.07	73.86	93.81	53.03	72.30	93.69	78.41	86.05
1.00×	0.005	0.60	97.33	1.000	21.72	0.141	0.662	89.25	98.19	70.83	93.53	52.72	72.19	93.29	77.49	85.39
2.00×	0.015	0.91	95.32	1.000	21.32	0.151	0.641	88.37	96.32	62.45	92.93	52.00	71.54	91.85	75.05	83.45

5.2 Static-Scene Track Results

Tab. 2 reports Camera Control and Scene Revisit together with the task-agnostic general metrics. The task scores diagnose *Control Adherence* and *Spatial Consistency*, whereas the general metrics characterize *Visual Quality*; reporting them separately reveals that these capabilities do not necessarily improve together.

Camera Control. Among camera-driven models, the strongest results come from methods that reconstruct 3D priors and reproject them to target views: NeoVerse [56] scores 97.33 and InSpatio-World [39] scores 85.94. ReCamMaster [3] and FantasyWorld [9], which encode camera poses as learned tokens or embeddings, obtain much lower Camera Control scores of 38.64 and 18.46 despite competitive General averages of 80.97 and 80.23. WorldPlay [38] is the strongest action-driven model at 92.74. Language-driven models are less precise, with Hailuo 2.3 [32] achieving the strongest score of 63.29, consistent with the difficulty of expressing ordered, complex viewpoint changes through natural-language instructions.

Scene Revisit. NeoVerse and InSpatio-World both achieve 1.000 Revisit Success and lead the camera-driven group with Scene Revisit scores of 89.25 and 85.90, respectively. Among action-driven models, WorldPlay performs best with 0.790 Revisit Success and a score of 72.51, while Hailuo 2.3 leads the language-driven group with 0.505 and 48.70. The remaining gap reflects failures either to return to the initial viewpoint or to recover its geometry, appearance, and content after the round trip.

5.3 Effect of the Input Translation Multiplier

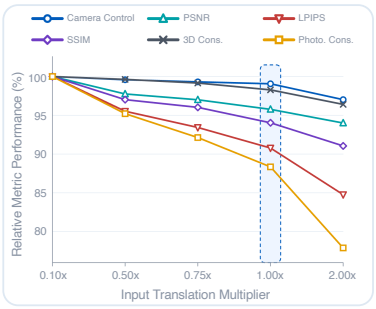


Figure 7 Effect of the input translation multiplier on NeoVerse.

We evaluate NeoVerse [56] under the same camera controls while varying only the translation multiplier of its input SE(3) trajectory. As the multiplier increases from 0.10× to 2.00×, the Camera Control score decreases from 98.25 to 95.32, the Scene Revisit score from 90.98 to 88.37, and the General average from 80.42 to 75.05; all five general metrics decline, with Photometric Consistency falling from 80.17 to 62.45. The results confirm that larger image-space displacements make both controlled generation and scene preservation more difficult. This ablation therefore motivates the pre-generation image-space displacement alignment in Sec. 4.2, which is applied to all reported camera-driven results.

5.4 Dynamic-Interaction Track Results

Tab. 4 reveals whether a model follows subject-motion control and whether it can react correctly.

Subject Control. The direct action interfaces provide more precise subject control: LingBot-World [40] scores 55.47 and WorldPlay [38] scores 49.75, compared with the best language-driven score of 37.28 from Veo 3.1 [12]. Even the action-driven results remain far from saturated, with failures often converting the requested subject motion into camera motion or leaving the scene static.

Terrain Interaction. Vidu Q3 [5] and Hailuo 2.3 [32] lead with 64.39 and 61.57, whereas the best action-driven score is 27.49. For action-driven models, the large drop from Subject Control to Terrain Interaction shows that horizontal control adherence does not guarantee vertical terrain adaptation.

Object Interaction. Veo 3.1 and Vidu Q3 lead with 75.96 and 71.59, while the best action-driven score is 33.75. Common failures leave the contacted object unchanged or allow the subject to pass through it.

Table 4 Dynamic-interaction track evaluation. Subject Control and five *World Reactivity* tasks are reported for compatible action- and language-driven models. **Task** averages the supported task scores, **General** averages the four general metrics, and **Overall** averages Task and General scores. Down ↓ and up ↑ arrows indicate that lower and higher values are better, respectively. Dashes mark unsupported tasks; the best and second-best results per paradigm are bold and underlined.

Model	Subject Control	Terrain Inter.	Object Inter.	Social Inter.	Physical Reaction	Goal Completion	Subject Cons.	Motion Smooth.	Aesth. Quality	Imag. Quality	Average		
											Task	General	Overall
Action-driven													
WorldPlay [38]	49.75	**27.49**	**33.75**	51.40	26.91	–	88.06	98.09	54.19	68.76	37.86	77.28	57.57
LingBot-World [40]	**55.47**	24.33	25.94	**60.37**	**33.43**	–	**94.98**	**98.89**	**60.86**	**71.69**	**39.91**	**81.61**	**60.76**
Language-driven													
Kling 2.5 [17]	28.40	35.95	27.70	66.80	31.99	48.25	**96.00**	**99.48**	56.86	71.83	39.85	**81.04**	60.45
Veo 3.1 [12]	**37.28**	44.71	**75.96**	**85.10**	61.76	85.30	92.07	99.12	**58.22**	72.65	**65.02**	80.52	**72.77**
Hailuo 2.3 [32]	36.49	61.57	67.01	72.45	**63.84**	78.86	93.25	99.31	57.74	72.47	63.37	80.69	72.03
Wan 2.6 I2V [44]	29.02	49.21	44.56	66.40	48.15	76.39	94.46	98.15	56.30	**74.72**	52.29	80.91	66.60
Seedance 1.5 [35]	32.51	53.83	37.91	72.09	47.60	76.24	92.14	98.87	56.72	70.84	53.36	79.64	66.50
Vidu Q3 [5]	27.67	**64.39**	71.59	81.91	61.23	78.26	92.78	98.76	57.24	73.25	64.18	80.51	72.35
HappyHorse 1.0 [1]	33.11	56.30	65.70	76.17	47.01	**85.33**	92.69	98.85	57.37	73.23	60.60	80.54	70.57

Social Interaction. Veo 3.1 achieves the highest score of 85.10, followed by Vidu Q3 at 81.91; the best action-driven score is 60.37. Failures typically leave nearby agents unresponsive or allow the controlled subject to move through them without avoidance or yielding.

Physical Reaction. Hailuo 2.3 leads with 63.84, followed by Veo 3.1 and Vidu Q3 at 61.76 and 61.23; the best action-driven score is 33.43. Action-driven generations often execute subject control while leaving unstable or contacted objects unchanged, exposing the gap between explicit control and inherent reactivity.

Goal Completion. HappyHorse 1.0 [1] and Veo 3.1 achieve the strongest results at 85.33 and 85.30. Kling 2.5 [17] scores only 48.25 despite having the highest General average among language-driven models, showing that visual quality does not guarantee grounded goal-directed execution.

5.5 Cross-Task Diagnostic Analysis

Fig. 8 consolidates the capability split across the three interfaces. Camera-driven models provide the strongest camera control and scene revisiting but do not support dynamic interaction. Action-driven models control designated subjects more precisely, yet this advantage does not consistently transfer to the scene-conditioned reactions induced by those controls. Language-driven models perform better on interaction and goal-directed tasks but follow composed camera and subject controls less faithfully. No model combines broad coverage with consistently strong performance, leaving current interfaces complementary but incomplete.

The split is also obscured by general visual metrics. For example, the language-driven General averages occupy a narrow range of 79.64–81.04, while their Task averages range from 39.85 to 65.02. Likewise, ReCamMaster and FantasyWorld retain strong General averages despite weak Camera Control scores. These gaps show that the four diagnostic levels capture distinct capabilities. In particular, strong *Visual Quality* or *Control Adherence* does not guarantee *World Reactivity*, motivating separate reporting of the four levels.

5.6 Human Alignment of Checklist Evaluation

Table 5 Human alignment of checklist evaluation. Spearman's ρ and PLCC (Pearson correlation) compare human and VLM checklist-satisfaction scores per task and across all 800 instances.

Task	Checklist Items	Spearman ρ	PLCC
Goal Completion	1,303	0.8960	0.9017
Physical Reaction	1,425	0.8838	0.8750
Object Interaction	1,527	0.8251	0.8541
Social Interaction	1,538	0.7019	0.7103
Overall	**5,793**	**0.8614**	**0.8583**

We validate the VLM judge on the four tasks evaluated using checklists. The validation set contains 800 evaluation instances and 5,793 checklist items, with 200 instances per task. Three human annotators independently label each item from the same 10 temporally ordered frames shown to the VLM, and the majority vote defines the binary reference label. For each instance, the human and VLM scores are computed as the respective fractions of satisfied checklist items. Tab. 5 reports Spearman's ρ and PLCC for each task separately and across all 800 evaluation instances combined. Across all 800 instances, Spearman's ρ is 0.8614 and PLCC is 0.8583, showing strong agreement between the VLM judge and human evaluation.

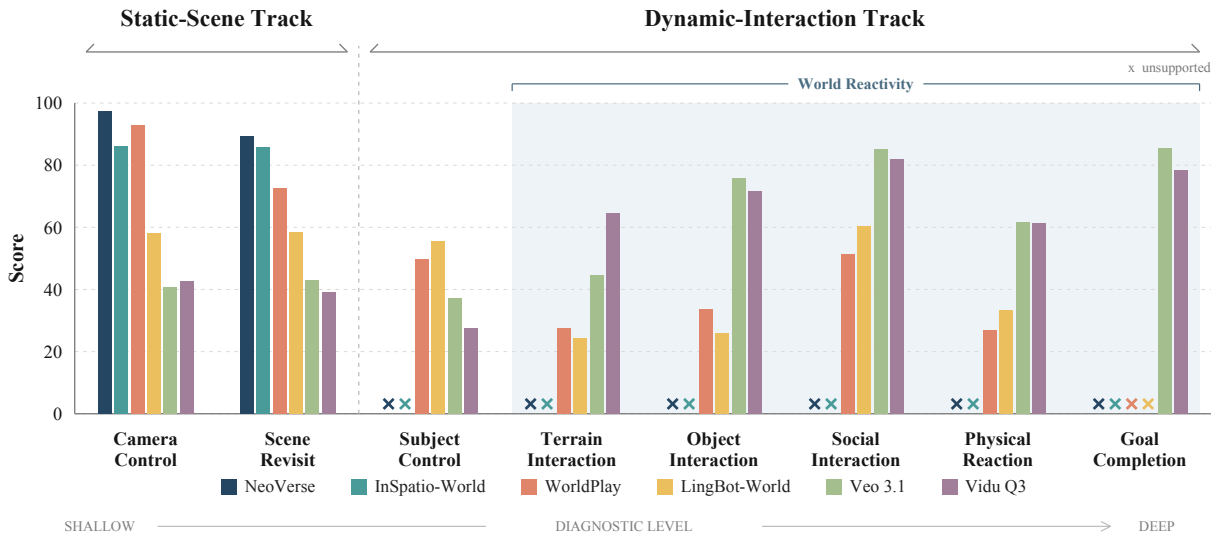


Figure 8 Task-level performance across evaluation tracks. Two models per interface are compared across all eight tasks, ordered from static-scene diagnostics through Subject Control to World Reactivity; crosses mark unsupported tasks.

5.7 Backend Stability with DA3 Reconstruction

To assess sensitivity to the reconstruction backend, we rerun the geometry-based metrics with Depth Anything 3 (DA3) [23] while keeping the benchmark inputs and model outputs fixed. Tabs. 6 and 7 report the corresponding results. Scores for the four tasks evaluated using checklists remain unchanged.

On the static-scene track, the mean absolute relative change in Overall score is 3.09% across all 20 models: 0.44% for camera-driven, 3.08% for action-driven, and 5.36% for language-driven models. Most variation is concentrated in Camera Control, while Scene Revisit and the general metrics change little. NeoVerse, WorldPlay, and Hailuo 2.3 remain the leading models in their respective groups; the camera- and action-driven rankings are fully preserved, with only closely matched language-driven models exchanging positions.

On the dynamic-interaction track, DA3 affects only Subject Control, Terrain Interaction, and their geometry-dependent aggregates. The mean absolute relative change in Overall score is 0.57%, the maximum change is 1.16%, and all within-paradigm rankings are preserved. Together, these small changes and stable rankings show that the main model comparisons do not depend on a particular reconstruction backend.

6 Conclusion

We presented WorldExam, a unified hierarchical benchmark for diagnosing video world models beyond visual quality and explicit instruction fulfillment. It distinguishes direct fulfillment of explicitly specified controls or targets from scene-conditioned reactions and detailed goal-directed execution that must be inferred from the initial scene. This distinction is instantiated through four diagnostic levels, eight tasks, and 1,474 test cases. Interface adaptation presents shared cases in the native formats of camera-, action-, and language-driven models, while the *static-scene* and *dynamic-interaction* tracks restrict evaluation to compatible interfaces rather than treating unsupported tasks as failures.

Evaluation of 20 representative models reveals a clear capability split. Camera-driven models provide the most precise camera control and scene revisiting; action-driven models control subjects more precisely but often leave terrain, objects, nearby agents, and physical processes unresponsive; and language-driven models perform better on interaction and goal-directed tasks but follow composed controls less faithfully. No evaluated model combines broad task coverage with consistently strong performance, showing that high visual quality and explicit instruction fulfillment do not guarantee *inherent reactivity*. Strong agreement between human

Table 6 Static-scene track evaluation with DA3. We recompute the static-scene metrics with DA3 while keeping benchmark inputs and model outputs fixed; aggregation and highlights follow Tab. 2.

Model	Camera Control			Scene Revisit					3D Cons.	Photo. Cons.	Temp. Flick.	Aesth. Quality	Imag. Quality	Average		
	T. Err.↓	R. Err.↓	Score↑	Success↑	PSNR↑	LPIPS↓	SSIM↑	Score↑						Task	General	Overall
Camera-driven																
TrajectoryCrafter [59]	0.13	1.10	81.97	**1.000**	19.06	0.240	0.544	82.99	95.63	62.83	92.56	52.18	67.03	82.48	74.05	78.27
ReCamMaster [3]	0.47	4.01	38.19	0.815	16.03	0.352	0.415	68.04	**99.18**	**81.97**	95.51	54.58	73.45	53.12	**80.94**	67.03
Voyager [13]	0.30	4.04	60.03	0.995	16.89	0.386	0.467	76.33	89.81	27.44	93.04	53.59	63.69	68.18	65.51	66.85
FantasyWorld [9]	1.05	7.76	18.24	0.555	15.39	0.323	0.373	55.51	98.79	75.60	**95.90**	**57.11**	**73.98**	36.87	80.28	58.58
NeoVerse [56]	**0.01**	**0.72**	**96.93**	**1.000**	**21.70**	**0.141**	**0.661**	**89.22**	97.53	70.83	93.53	52.72	72.19	**93.07**	77.36	**85.22**
InSpatio-World (1.3B) [39]	0.08	1.21	86.91	**1.000**	20.77	0.225	0.605	85.84	97.79	64.88	93.68	53.77	73.60	86.38	76.74	81.56
Action-driven																
Hunyuan-GameCraft [20]	0.13	11.92	39.17	0.375	13.37	0.542	0.369	41.26	94.00	53.93	93.91	54.98	71.70	40.21	73.70	56.96
Astra [65]	0.71	7.97	15.97	0.360	12.18	0.600	0.309	37.89	97.09	78.42	96.37	51.71	71.78	26.93	79.07	53.00
WorldPlay [38]	**0.08**	**1.06**	**88.31**	0.780	**18.69**	**0.269**	**0.544**	**72.52**	98.55	81.06	95.90	53.92	73.31	**80.41**	80.55	**80.48**
Yume 1.5 [30]	0.32	3.14	59.69	0.270	12.37	0.614	0.359	33.40	97.74	67.53	95.18	53.42	75.27	46.55	77.83	62.19
LingBot-World [40]	0.19	6.28	54.43	0.620	15.63	0.340	0.409	59.17	99.42	81.52	**96.53**	**59.87**	75.01	56.80	82.47	69.64
Infinite-World [50]	0.25	2.38	64.49	0.655	14.30	0.389	0.345	57.76	**99.87**	**92.09**	95.99	56.27	**76.40**	61.12	**84.12**	72.62
Matrix-Game 3.0 [46]	0.17	4.27	68.93	**0.860**	13.83	0.486	0.372	64.23	95.61	30.14	93.39	48.80	73.39	66.58	68.27	67.43
Language-driven																
Kling 2.5 [17]	0.31	6.93	43.37	0.315	13.50	0.535	0.380	38.13	**99.78**	**87.87**	**97.57**	56.28	75.70	40.75	**83.44**	62.10
Veo 3.1 [12]	0.89	8.20	21.27	0.447	12.30	0.595	0.345	43.03	98.73	73.61	95.42	55.49	76.57	32.15	79.96	56.06
Hailuo 2.3 [32]	**0.28**	**5.18**	**54.61**	**0.510**	**13.56**	**0.518**	**0.388**	**49.00**	98.94	79.88	95.32	**56.30**	75.52	**51.81**	81.19	**66.50**
Wan 2.6 I2V [44]	0.38	6.12	46.50	0.485	13.29	0.527	0.355	46.88	99.27	69.43	94.13	53.93	**77.53**	46.69	78.86	62.78
Seedance 1.5 [35]	0.78	7.78	28.47	0.370	12.29	0.601	0.343	39.01	95.96	56.60	94.82	54.45	74.89	33.74	75.34	54.54
Vidu Q3 [5]	0.49	8.48	30.69	0.360	12.67	0.582	0.352	39.14	99.13	71.27	95.01	55.38	76.73	34.92	79.50	57.21
HappyHorse 1.0 [1]	0.40	5.55	45.66	0.415	12.45	0.564	0.351	42.16	99.52	74.17	95.02	55.57	76.99	43.91	80.25	62.08

Table 7 Dynamic-interaction track evaluation with DA3. We recompute Subject Control, Terrain Interaction, and affected aggregates with DA3; tasks evaluated using checklists remain unchanged, and aggregation and highlights follow Tab. 4.

Model	Subject Control	Terrain Inter.	Object Inter.	Social Inter.	Physical Reaction	Goal Completion	Subject Cons.	Motion Smooth.	Aesth. Quality	Imag. Quality	Average		
											Task	General	Overall
Action-driven													
WorldPlay [38]	47.51	**23.05**	**33.75**	51.40	26.91	–	88.06	98.09	54.19	68.76	36.52	77.28	56.90
LingBot-World [40]	**53.43**	21.63	25.94	**60.37**	**33.43**	–	**94.98**	**98.89**	**60.87**	**71.70**	**38.96**	**81.61**	**60.29**
Language-driven													
Kling 2.5 [17]	26.78	37.50	27.70	66.80	31.99	48.25	**96.00**	**99.49**	56.86	71.83	39.84	**81.05**	60.45
Veo 3.1 [12]	35.12	42.19	**75.96**	**85.10**	61.76	85.30	92.07	99.12	**58.21**	72.65	**64.24**	80.51	**72.38**
Hailuo 2.3 [32]	**35.16**	59.34	67.01	72.45	**63.84**	78.86	93.26	99.31	57.74	72.47	62.78	80.70	71.74
Wan 2.6 I2V [44]	26.95	48.55	44.56	66.40	48.15	76.39	94.46	98.15	56.30	**74.71**	51.83	80.91	66.37
Seedance 1.5 [35]	29.41	48.28	37.91	72.09	47.60	76.24	92.14	98.87	56.71	70.84	51.92	79.64	65.78
Vidu Q3 [5]	26.15	**62.56**	71.59	81.91	61.23	78.26	92.78	98.76	57.24	73.25	63.62	80.51	72.07
HappyHorse 1.0 [1]	31.62	53.81	65.70	76.17	47.01	**85.33**	92.69	98.86	57.37	73.23	59.94	80.54	70.24

and VLM checklist scores, together with stable model rankings under an alternative reconstruction backend, supports the reliability of these findings.

The current scope is bounded by the capabilities of available model interfaces. Dynamic-interaction evaluation requires reliable third-person subject control, and Goal Completion remains limited to language-driven models. Moreover, the metrics assess observable end-to-end behavior rather than determining where reasoning occurs or establishing that the video generator itself has learned an internal causal representation. For closed-source commercial systems, proprietary prompt enhancement may contribute to scene grounding and execution planning. Future extensions to broader interfaces, longer-horizon interactions, and more intervention-based settings would provide stronger tests of persistent world understanding. Within its current scope, WorldExam identifies where controllable video models succeed, where their generated worlds remain unresponsive, and which capabilities must be developed jointly.

Appendix

A Gallery

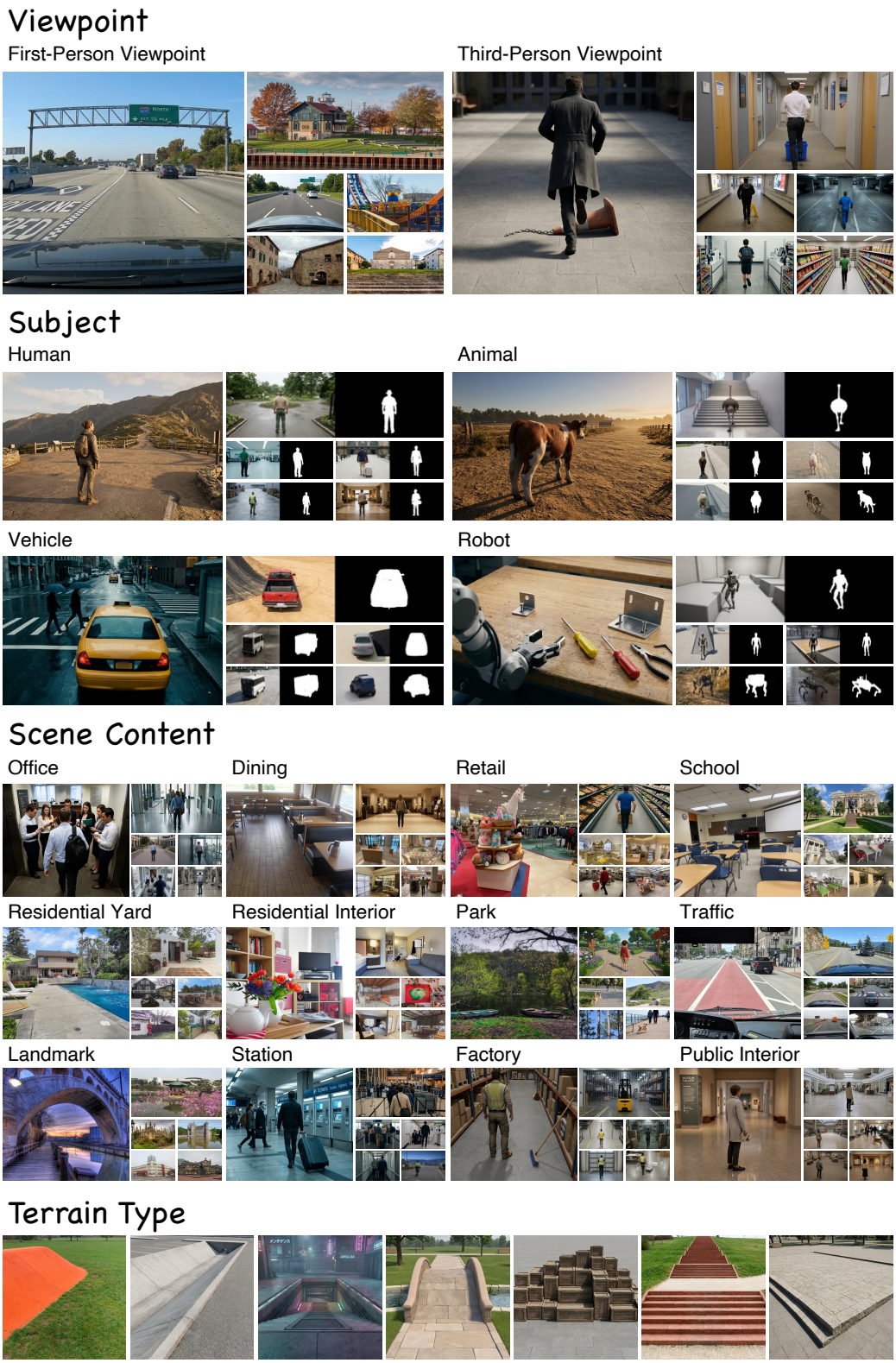


Figure 9 WorldExam Gallery. Representative test cases span first- and third-person viewpoints; human, animal, vehicle, and robot subjects; diverse indoor and outdoor scene content; and a range of terrain types.

B Per-Model Inference Settings

To preserve each model's native performance, we use its default resolution, video length, and other inference settings whenever applicable. For streaming models with flexible generation lengths, we constrain the output to 100–200 frames to prevent quality drift in excessively long generations from biasing the evaluation. Tab. 8 reports the resulting resolution and frame count for each model.

Eligibility for the dynamic-interaction track requires reliable control of a visible third-person subject. Among action-driven models, only WorldPlay [38] and LingBot-World [40] satisfy this requirement; the other five either do not support third-person subject control or cannot provide it reliably. All evaluated language-driven models accept third-person subject-motion prompts, whereas camera-driven interfaces control only the camera. Goal Completion is language-only within the dynamic-interaction track.

Table 8 Per-model inference settings and eligibility for the dynamic-interaction track. Resolution is the center-cropped input size in width×height order, and Frames reports the evaluated video length. *TPV* and *FPV* denote third- and first-person viewpoints, respectively. ✓ denotes eligibility for the dynamic-interaction track; ✗ denotes FPV-only control or unreliable TPV subject control. † denotes unreliable TPV subject control. A dash denotes that the track is not applicable because the model controls only the camera.

Model	Backend	Resolution	Frames	Dynamic Track
Camera-driven				
TrajectoryCrafter [59]	Local	672×384	49	–
ReCamMaster [3]	Local	832×480	81	–
Voyager [13]	Local	768×512	49	–
FantasyWorld [9]	Local	592×336	81	–
NeoVerse [56]	Local	560×336	81	–
InSpatio-World (1.3B) [39]	Local	832×480	81	–
Action-driven				
Hunyuan-GameCraft [20]	Local	1216×704	132	✗ TPV†
Astra [65]	Local	832×480	161	✗ FPV only
WorldPlay [38]	Local	832×480	125	✓ TPV
Yume 1.5 [30]	Local	1280×704	145	✗ FPV only
LingBot-World [40]	Local	832×464	161	✓ TPV
Infinite-World [50]	Local	896×448	161	✗ FPV only
Matrix-Game 3.0 [46]	Local	1280×704	177	✗ FPV only
Language-driven				
Kling 2.5 [17]	API	1280×720	153	✓ TPV prompt
Veo 3.1 [12]	API	1280×720	192	✓ TPV prompt
Hailuo 2.3 [32]	API	1024×768	141	✓ TPV prompt
Wan 2.6 I2V [44]	API	1280×720	150	✓ TPV prompt
Seedance 1.5 [35]	API	1280×720	97	✓ TPV prompt
Vidu Q3 [5]	API	1280×720	121	✓ TPV prompt
HappyHorse 1.0 [1]	API	1280×720	123	✓ TPV prompt

C Qualitative Examples of the Eight Tasks

The examples below instantiate all eight tasks in a shared visual format. The highlighted image is the initial frame, followed by four temporally ordered generated frames. The bottom panel shows the shared control intent or high-level goal. For tasks evaluated using checklists, the checklist is shown only to explain the evaluation protocol and is withheld from the input.

C.1 Camera Control

Camera Control measures whether an ordered composition of camera motions follows the prescribed directions and temporal order without unintended drift. In this case, the camera tilts down, pans left, and moves left.

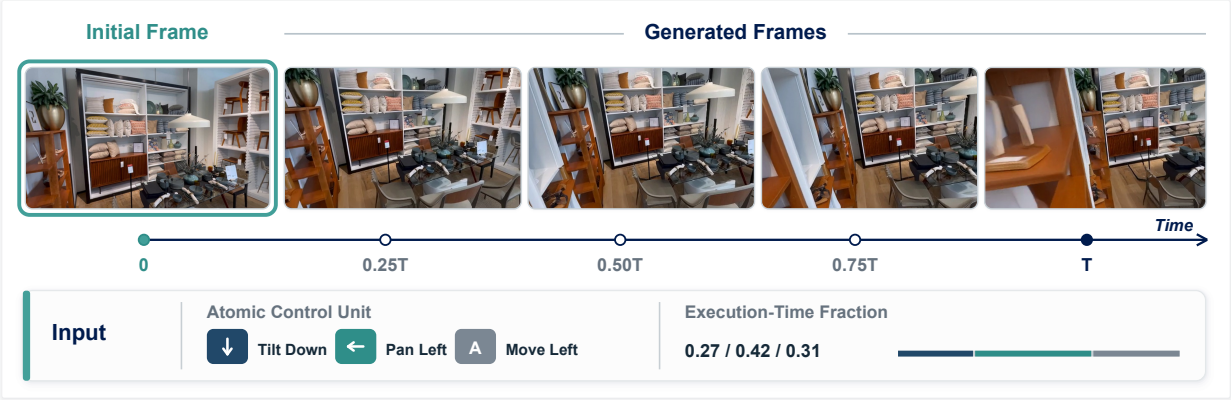


Figure 10 Camera Control. The three atomic camera controls occupy 0.27, 0.42, and 0.31 of the video, respectively.

C.2 Subject Control

Subject Control applies atomic controls to a designated third-person subject. Here, the subject should move forward, right, and left in order while remaining visually identifiable.

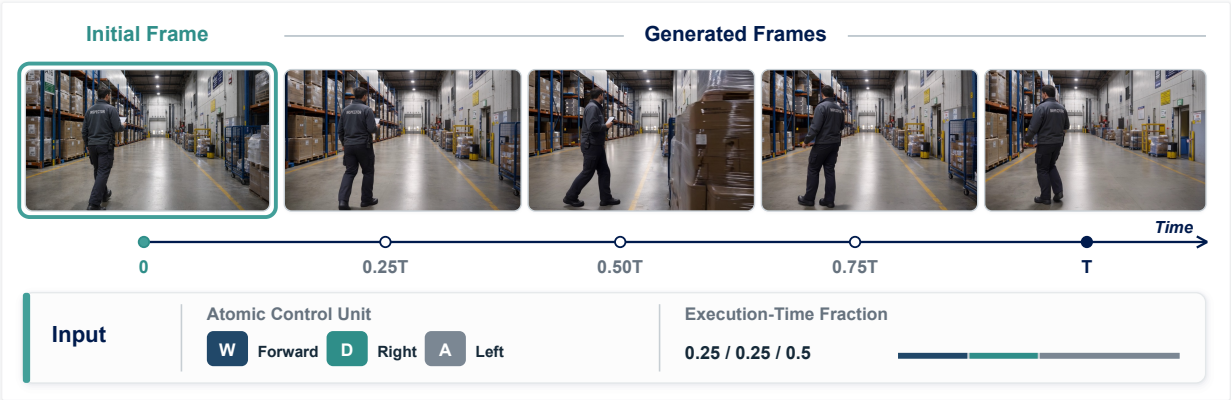


Figure 11 Subject Control. The Forward, Right, and Left controls occupy 0.25, 0.25, and 0.50 of the video.

C.3 Scene Revisit

Scene Revisit couples a round-trip camera motion with a spatial-memory requirement: the camera should return to the initial viewpoint while preserving the scene.

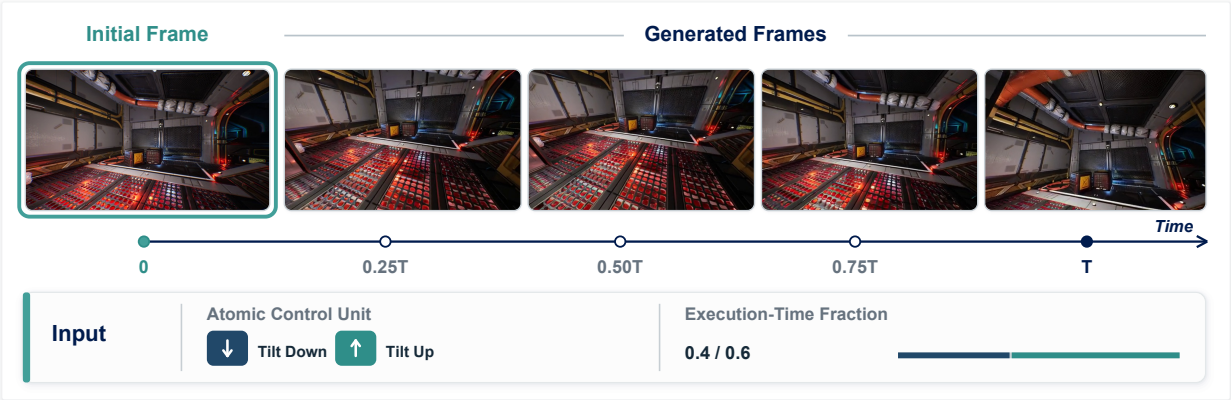


Figure 12 Scene Revisit. The camera first tilts down and then tilts up; success requires both execution of the motion and recovery of a consistent revisited view.

C.4 Terrain Interaction

Terrain Interaction specifies only horizontal subject motion. The model must infer the vertical adaptation required by the visible terrain—in this case, traversing the stairs while continuing forward.

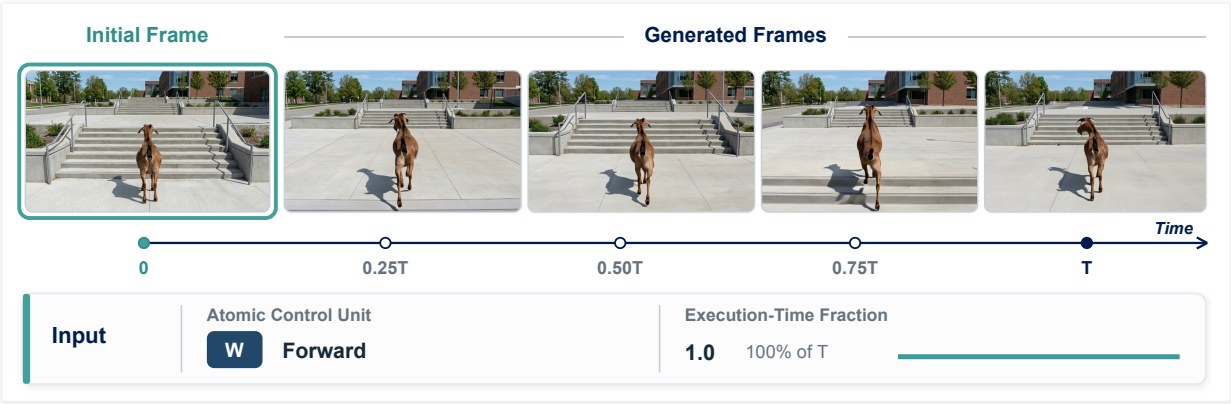


Figure 13 Terrain Interaction. The Forward control is active for the full video, while the stair-climbing response is implied by the scene rather than stated in the input.

C.5 Object Interaction

Object Interaction tests whether contact with a designated target causes a response consistent with the target's physical type. Here, the worker pushes a bus cart forward into a lightweight sign.

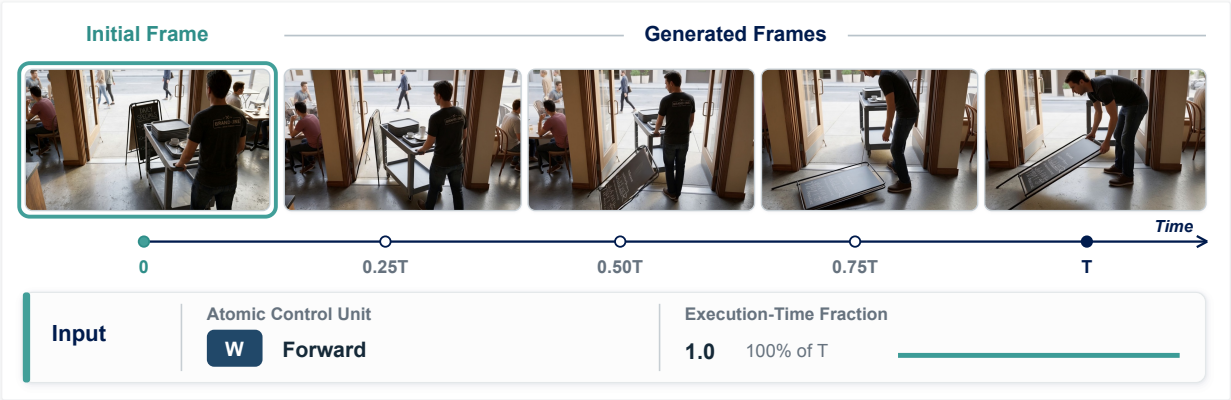


Figure 14 Object Interaction. Only the Forward control is specified; the sign's contact response is withheld from the model input and assessed with the checklist below.

Case-Specific Evaluation Checklist

Evaluation only; not part of the model-facing input.

- ❑ Does the worker keep the bus cart rolling straight forward on the same approach line into the sign without steering away before impact?
- ❑ Does the bus cart make the first contact with the sign, specifically low on the sign's lower frame, panel edge, or legs, rather than the worker's body hitting it directly?
- ❑ After contact, does the A-frame sign behave like a light freestanding hinged sign by tipping, folding, sliding, or getting shoved aside?
- ❑ Does the sign remain visibly unattached and mobile, rather than acting as if fixed rigidly to the floor or doorway?
- ❑ Does the sign's motion follow the cart's low forward push, with the base and legs reacting first instead of the top moving in an implausible independent way?
- ❑ Does the worker stay behind the cart with a continued forward pushing posture through the moment of impact?
- ❑ If the cart continues through the doorway area, is that continuation enabled by the sign being displaced or collapsing aside rather than unrealistically blocking the cart like a solid barrier?
- ❑ Does the continuation keep the cart–sign contact zone and the sign's passive response visible enough to judge the impact clearly?

C.6 Social Interaction

Social Interaction evaluates whether nearby agents respond plausibly when a controlled subject enters their social or safety space. Here, a sedan enters a storefront crossing with two pedestrians and a shopping cart.

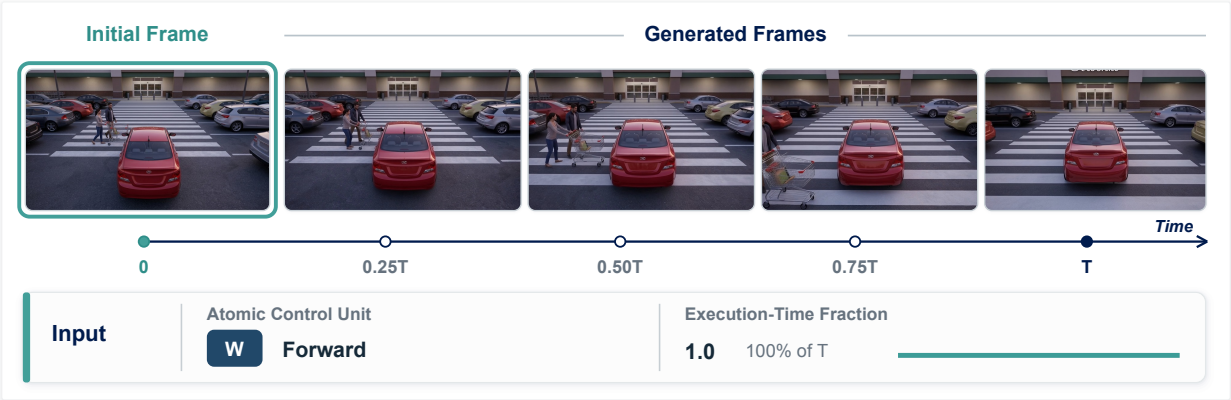


Figure 15 Social Interaction. Only the car's Forward control is specified; the checklist evaluates the pedestrians' response.

Social Interaction Evaluation Checklist

Evaluation only; not part of the model-facing input.

- ❑ Does the red sedan continue inching forward into the storefront crossing?
- ❑ Do the cart pusher and companion remain visible enough to judge how they pass the car nose?
- ❑ Does the cart path or walking pace visibly change near the sedan's protruding front end?
- ❑ Is the cart interaction focused on the crossing area directly in front of the car?
- ❑ Does the continuation stay at ordinary parking-lot speed and behavior?
- ❑ Does the sedan remain the dominant source of social pressure in the scene?
- ❑ Do the pedestrians guiding the shopping cart respond plausibly as they move around the vehicle nose?
- ❑ Does the clip avoid unrelated dominant events such as another car cutting through the scene?

C.7 Physical Reaction

Physical Reaction evaluates the temporal development of a physical process, not merely whether contact occurs. Here, the woman remains stationary while a towel-loaded laundry basket, whose center of mass overhangs the washer edge, begins to tip and fall.

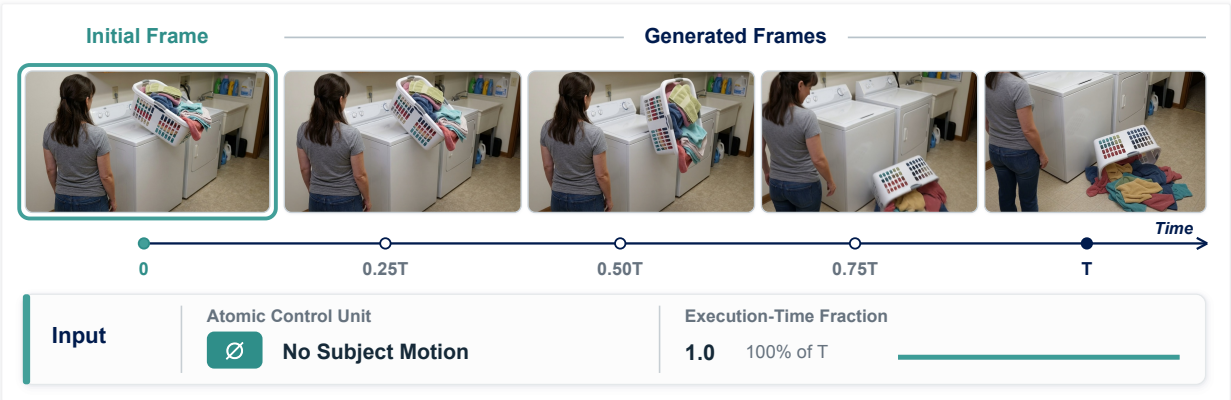


Figure 16 Physical Reaction. The ∅ (stop) control specifies no subject motion; the basket's rotation about the support edge and subsequent fall should unfold autonomously.

Physical Reaction Evaluation Checklist

Evaluation only; not part of the model-facing input.

- ❑ Does the woman remain stationary throughout the short continuation?
- ❑ Does the basket continue rotating outward and downward from its already tilted position?

- ❑ Does the basket fall off the washing machine rather than sliding back to a fully supported position on top?
- ❑ Does the motion begin as a tip about the washer edge or remaining contact point, consistent with the overhanging weight pulling it over?
- ❑ Does the overhanging towel load move with the basket and contribute to the same outward fall direction?
- ❑ Does the basket fall without any new touch, grab, or bump from the woman?
- ❑ After leaving support, do the basket and laundry move downward into the open space in front of the machine rather than floating or reversing upward?

C.8 Goal Completion

Goal Completion provides a high-level goal instead of an atomic control sequence. The model must ground the goal in the initial scene and generate coherent execution steps toward the desired target.

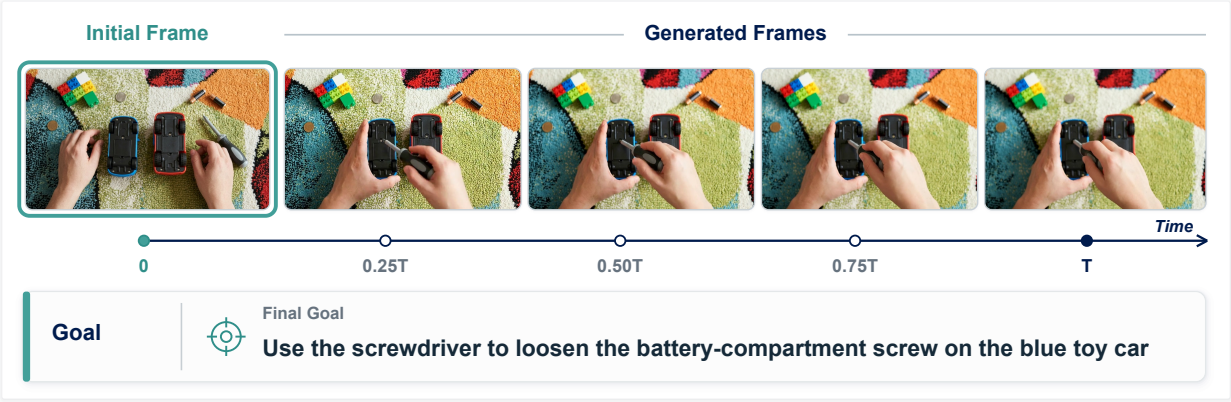


Figure 17 Goal Completion. The goal is to loosen the blue toy car's battery-compartment screw with the screwdriver.

Goal Completion Evaluation Checklist

Evaluation only; not part of the model-facing input.

- ❑ Does the person pick up and use the visible screwdriver?
- ❑ Does the action target the blue toy car rather than the red toy vehicle?
- ❑ Does the screwdriver tip align with the blue car's battery-compartment screw?
- ❑ Does the continuation show a clear screw-loosening motion on the blue car?
- ❑ Does the red toy vehicle remain unused throughout the task?
- ❑ Does the person avoid using the nearby coin as a tool?
- ❑ Does the blue car's battery cover end visibly loosened or ready to open?

References

[1] Alibaba Group. HappyHorse 1.0 I2V, 2026. URL https://www.alibabacloud.com/help/en/model-studio/happyhorse-image-to-video-api-reference.

[2] Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. Act-bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*, 2024.

[3] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *ICCV*, pages 14834–14844. IEEE, 2025.

[4] 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 *ICLR*, volume 2025, pages 102075–102121, 2025.

[5] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

[7] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

[8] Yubin Chen, Xuyang Guo, Zhenmei Shi, Zhao Song, and Jiahao Zhang. T2vworldbench: A benchmark for evaluating world knowledge in text-to-video generation. In *WACV*, pages 6474–6485. IEEE, 2026.

[9] Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *ICLR*, volume 2026, pages 103603–103622, 2026.

[10] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *ICCV*, pages 27713–27724, 2025.

[11] Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, et al. iworld-bench: A benchmark for interactive world models with a unified action generation framework. *arXiv preprint arXiv:2605.03941*, 2026.

[12] Google DeepMind. Veo 3 technical report. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

[13] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM TOG*, 44(6):1–15, 2025.

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818. IEEE, 2024.

[15] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *TPAMI*, 2025.

[16] Feng Jiang, Yang Chen, Kyle Xu, Yuchen Liu, Haifeng Wang, Zhenhao Shen, Jasper Lu, Shengze Huang, Yuanfei Wang, Chen Xie, et al. Robowm-bench: A benchmark for evaluating world models in robotic manipulation. *arXiv preprint arXiv:2604.19092*, 2026.

[17] Kuaishou Technology. Kling AI. https://klingai.com, 2025. URL https://klingai.com.

[18] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *NeurIPS*, 38, 2026.

[19] Huiqiong Li, Jiayu Wang, Zhiting Mei, Anirudha Majumdar, Jingjing Chen, and Bin Zhu. Robotrustbench: Benchmarking the trustworthiness of video world models for robotic manipulation. *arXiv preprint arXiv:2606.01600*, 2026.

[20] Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025.

[21] Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025.

[22] Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Yu Yang, Ziqi Huang, Wei Yin, Jialong Zuo, Yixuan Hu, Dekai Zhu, et al. Worldlens: Full-spectrum evaluations of driving world models in real world. In *CVPR*, pages 36385–36399, 2026.

[23] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[24] Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y Lu, Enfu Nan, Hokin Deng, et al. Phyground: Benchmarking physical reasoning in generative world models. *arXiv preprint arXiv:2605.10806*, 2026.

[25] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169. IEEE, 2024.

[26] Mingxin Liu, Shuran Ma, Shibei Meng, Xiangyu Zhao, Zicheng Zhang, Shaofeng Zhang, Zhihang Zhong, Peixian Chen, Haoyu Cao, Xing Sun, et al. Rise-video: Can video generators decode implicit world rules? *arXiv preprint arXiv:2602.05986*, 2026.

[27] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, pages 22139–22149. IEEE, 2024.

[28] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *NeurIPS*, 36:62352–62387, 2023.

[29] Zeyu Liu, Zhangzhe Zhu, Yang Zhang, Chenyou Fan, Chenjia Bai, and Xuelong Li. Kinebench: Benchmarking embodied world models via idm-free kinematic grounding. *arXiv preprint arXiv:2607.19876*, 2026.

[30] Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *CVPR*, pages 7752–7761, 2026.

[31] 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. *arXiv preprint arXiv:2410.05363*, 2024.

[32] MiniMax. Hailuo AI Video. https://hailuoai.video/, 2024. URL https://hailuoai.video/.

[33] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, volume 2025, pages 28085–28128, 2025.

[35] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[36] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[37] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *CVPR*, pages 8406–8416. IEEE, 2025.

[38] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

[39] InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling. *arXiv preprint arXiv:2604.07209*, 2026.

[40] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

[41] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, pages 1110–1121. IEEE, 2017.

[42] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal processing*, 167:107299, 2020.

[43] Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Pranay Boreddy, Shruti Satya Narayana, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. Worldbench: Disambiguating physics for diagnostic evaluation of world models. *arXiv preprint arXiv:2601.21282*, 2026.

[44] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[45] Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω. *arXiv preprint arXiv:2605.15195*, 2026.

[46] Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

[47] Jiaxin Wu, Yihao Pi, Yinling Zhang, Yuheng Li, and Xueyan Zou. Quantitative video world model evaluation for geometric-consistency. *arXiv preprint arXiv:2605.15185*, 2026.

[48] Keming Wu, Yijing Cui, Wenhan Xue, Qijie Wang, Xuan Luo, Zhiyuan Feng, Zuhao Yang, Sudong Wang, Sicong Jiang, Haowei Zhu, et al. Worldreasonbench: Human-aligned stress testing of video generators as future world-state predictors. *arXiv preprint arXiv:2605.10434*, 2026.

[49] Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, et al. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026.

[50] Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026.

[51] Ting-Bing Xu, Jiacheng Sui, Zhe Gao, Kewei Shi, Wenjin Yang, Zhicheng Liu, Zhaoxu Sun, Mingchao Sun, Hongyu Pan, Fan Jiang, et al. Worldroambench: An open-world benchmark for long-horizon stability of interactive world models. *arXiv preprint arXiv:2606.31672*, 2026.

[52] Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. Worldmark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026.

[53] Haotian Xue, Yipu Chen, Liqian Ma, Zelin Zhao, Lama Moukheiber, Yuchen Zhu, and Yongxin Chen. Acwm-phys: Investigating generalized physical interaction in action-conditioned video world models. *arXiv preprint arXiv:2605.08567*, 2026.

[54] Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

[55] Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. Mirabench: Evaluating action-conditioned reliability in robotic world models. *arXiv preprint arXiv:2605.29360*, 2026.

[56] Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

[57] Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

[58] Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

[59] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, pages 100–111. IEEE, 2025.

[60] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

[61] Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. World-in-world: World models in a closed-loop world. In *ICLR*, volume 2026, pages 55660–55699, 2026.

[62] Yuke Zhao, Wangbo Zhao, Weijie Wang, Zeyu Zhang, Dakai An, Akide Liu, Yinghao Yu, Jiasheng Tang, Fan Wang, Wei Wang, et al. Worldolympiad: Can your world model survive a triathlon? *arXiv preprint arXiv:2606.11129*, 2026.

[63] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[64] Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven Waslander. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. In *ICLR*, volume 2026, pages 103502–103524, 2026.

[65] Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *ICLR*, volume 2026, pages 79167–79184, 2026.