WORLDEXAM: BENCHMARKING WORLD MODELS FROM APPARENT APPEARANCE TO INHERENT REACTIVITY

Anonymous authors
Paper under double-blind review



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.

ABSTRACT

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 with expected consequences left unstated and goal-directed behaviors with execution steps unspecified. Evaluation of 23 models shows that neither high visual quality nor strong control adherence guarantees inherent reactivity, while models with stronger reactivity can exhibit less reliable control. These findings suggest that progress requires combining these strengths, so that models both follow controls and generate plausible consequences implied by the scene.

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

Table 1: **Comparison with representative world-model benchmarks.** C / A / L denote camera-, action-, and language-driven model paradigms. † indicates that the instruction specifies the expected interaction consequence; the corresponding WorldExam tasks leave them 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 (Duan et al., 2025)	C / L	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	3,000	20
MIND (Ye et al., 2026)	A	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	250	2
Omni-WorldBench (Wu et al., 2026b)	C / L	✓	✓	✓	✓	✓	✗	✓†	✗	✓†	✗	1,068	18
WorldMark (Xu et al., 2026b)	C / A / L	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	500	10
iWorld-Bench (Fang et al., 2026)	C / A / L	✓	✗	✓	✗	✓	✗	✗	✗	✗	✗	4,900	14
WBench (Ying et al., 2026)	C / A / L	✓	✓	✓	✓	✓	✗	✓	✗	✓	✗	289	20
WorldOlympiad (Zhao et al., 2026)	A / L	✓	✗	✓	✗	✗	✗	✓†	✗	✓†	✗	1,000	8
WorldRoamBench (Xu et al., 2026a)	A	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	1,006	12
WorldExam (Ours)	C / A / L	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**1,474**	**23**

1 INTRODUCTION

Controllable video generation models are increasingly being developed as world models rather than standalone clip generators (Brooks et al., 2024; Bruce et al., 2024; Yang et al., 2023; 2026; Sun et al., 2025b; Bao et al., 2024; Google DeepMind, 2025). They are expected to predict future visual states from an initial observation and control instructions. Evaluating them in this role extends beyond the *apparent appearance* of generated videos to the *inherent reactivity* of the worlds they depict. When a subject moves onto stairs, its motion should adapt to the terrain; when it approaches an obstacle, the scene should exhibit contact, avoidance, or blockage; when it enters another agent's personal space, that agent should respond plausibly. We call the ability to infer and generate plausible consequences implied by the scene but not explicitly requested by the control *inherent reactivity*.

Recent benchmarks have advanced world-model evaluation beyond perceptual quality (Huang et al., 2024) to structured layout control (Duan et al., 2025), interactive control (Ying et al., 2026; Xu et al., 2026a), prompt-specified interaction effects (Wu et al., 2026b; Zhao et al., 2026), and embodied-AI and autonomous-driving applications (Shang et al., 2026; Liang et al., 2026). Complementary benchmarks probe implicit rules, future-state reasoning, and law-specific physical consistency in specialized settings (Liu et al., 2026; Wu et al., 2026a; Upadhyay et al., 2026; Lin et al., 2026). Table 1 compares representative benchmarks. Yet these advances leave a gap in systematically diagnosing inherent reactivity across various domains, including terrain, objects, other agents, and physical reactions, with expected consequences left unstated in the inputs.

We introduce **WorldExam**, a hierarchical diagnostic benchmark comprising 1,474 cases across eight tasks and four diagnostic levels: *Visual Quality*, *Control Adherence*, *Spatial Consistency*, and *World Reactivity* (Figure 1). The first three levels assess foundational capabilities, while *World Reactivity* provides a dedicated five-task suite spanning terrain, object, and social interactions, physical reactions, and goal completion. The first four reaction tasks pair an initial image with controls while leaving expected consequences unstated; Goal Completion states a target but leaves its execution steps open. WorldExam adapts shared atomic controls to camera-, action-, and language-driven interfaces, enabling unified evaluation for different model paradigms. The static-scene track evaluates Camera Control and Scene Revisit across all three interfaces; the dynamic-interaction track covers Subject Control and five World Reactivity tasks for compatible action- and language-driven models.

Evaluation of 23 models shows that strong control adherence does not consistently extend to appropriate scene-conditioned reactions, while stronger reactivity can coexist with less reliable control. Visual quality alone does not fully reflect world-model capabilities: models producing similarly high-quality videos can differ substantially in control adherence and inherent reactivity. Our evaluation combines geometry-based metrics with VLM checklist judgments. Human agreement and consistent cross-task analysis across reconstruction backends support evaluation reliability.

Our contributions to video world-model evaluation are as follows.

- We systematically evaluate inherent reactivity by probing interactions with various elements of the world, including terrain, objects, and other agents, as well as physical reactions, with expected consequences left unstated in the inputs.

- We introduce WorldExam, comprising 1,474 cases across eight tasks and four diagnostic levels, with shared controls adapted to camera-, action-, and language-driven interfaces and image-space displacement calibration for fair comparison of camera-driven models.
- We evaluate 23 models and identify distinct strengths in visual quality, control adherence, spatial consistency, and inherent reactivity, with evaluation reliability supported by human agreement and reconstruction-backend checks.

We will release the benchmark data and toolkit for reproducible evaluation of video world models.

2 Related Work

2.1 Video World Models

Video world models can be grouped by their primary control interfaces into camera-, action-, and language-driven paradigms. Camera-driven systems use learned trajectory encodings or reconstruct and reproject 3D priors (Yu et al., 2025; Bai et al., 2025; Huang et al., 2025a; Dai et al., 2026; Yang et al., 2026; Team et al., 2026a). Action-driven systems generate future frames from discrete, keyboard-like control sequences (Li et al., 2025; Zhu et al., 2026b; Sun et al., 2025b; Mao et al., 2026; Team et al., 2026b; Wu et al., 2026c; Wang et al., 2026b). Language-driven systems condition generation on text or image–text prompts specifying motion, interactions, and goals (Kuaishou Technology, 2025; Google DeepMind, 2025; MiniMax, 2024; Wan et al., 2025; Seedance et al., 2025; Bao et al., 2024; Alibaba Group, 2026). To accommodate these differences, WorldExam adapts shared atomic controls to each model's native interface and evaluates compatible tasks.

2.2 Video World Model Benchmarks

General video benchmarks assess visual quality, compositionality, and physical plausibility (Huang et al., 2024; 2025b; Zheng et al., 2025; Sun et al., 2025a; Bansal et al., 2025; Meng et al., 2024). For controllable world generation, WorldScore (Duan et al., 2025) evaluates camera-based layout control and geometric consistency, while MIND (Ye et al., 2026) probes action control and revisit consistency. WorldMark (Xu et al., 2026b) and iWorld-Bench (Fang et al., 2026) support cross-model comparison through unified action representations. Omni-WorldBench (Wu et al., 2026b) evaluates prompt-specified interaction outcomes and causal transitions, while WorldOlympiad (Zhao et al., 2026) probes long-horizon interaction and physics. RISE (Liu et al., 2026), WorldReasonBench (Wu et al., 2026a), WorldBench (Upadhyay et al., 2026), and PhyGround (Lin et al., 2026) examine implicit rules, future-state reasoning, and physical consistency. Although WBench (Ying et al., 2026) and WorldRoamBench (Xu et al., 2026a) also assess scene-conditioned physical consequences, WorldExam adopts a broader view of world reactivity, with task coverage compared in Table 1, through systematic diagnosis of terrain, object, and social interactions and physical reactions. Its reaction tasks leave expected consequences unstated in the control inputs and include dedicated social-response assessment and geometric terrain evaluation. The suite is complemented by a higher-level Goal Completion task that tests execution steps toward a stated target.

3 Benchmark Design

WorldExam evaluates how a generated world follows controls and produces the reactions implied by its initial scene. We formulate a video world model as $f : \mathcal{I} \times \mathcal{C} \to \mathcal{V}$, mapping an initial image and a native control input to a video. To accommodate different interfaces, we represent control intent through shared atomic units: W/S/A/D for translation, arrow keys for camera tilt or pan, and $\varnothing$ for stopping. Interface adaptation maps ordered compositions of these units to $\mathrm{SE}(3)$ camera trajectories, discrete action sequences, or natural-language prompts.

3.1 Toward World Reactivity: A Four-Level Diagnostic Hierarchy

Figure 1 organizes evaluation into four diagnostic levels of increasing scope. *Visual Quality* assesses apparent appearance, including perceptual plausibility and temporal stability. *Control Adherence* checks whether the camera or subject follows the input controls, while *Spatial Consistency* checks whether the world is preserved when the camera revisits an earlier viewpoint.

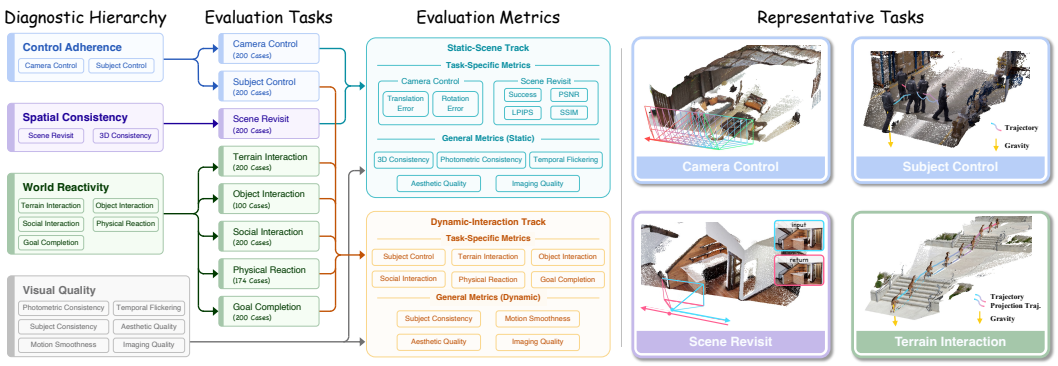


Figure 2: **WorldExam taxonomy, tracks, and metrics.** Eight tasks connect diagnostic levels to track-specific evaluation and general metrics, with geometry-based examples.

World Reactivity examines scene-conditioned reactions and goal-directed behaviors. Reaction inputs specify an initiating behavior but leave consequences unstated: moving forward may require climbing a step, displacing an object, or eliciting another agent's response, with these consequences defining the reaction criteria. Goal-directed evaluation instead specifies a target while leaving the detailed execution steps for the model to determine.

3.2 From Diagnostic Levels to Evaluation Tasks

Visual Quality uses task-agnostic metrics, while the other three levels are instantiated as eight evaluation tasks, as illustrated in Figure 2.

Control Adherence and Spatial Consistency. **Camera Control** composes one to three translation or rotation atomic control units and assigns an execution-time fraction to each. **Subject Control** applies the same construction to a designated third-person subject using W/S/A/D. **Scene Revisit** pairs an outgoing camera control with its inverse, such as moving left and then right, testing whether the camera returns to its initial viewpoint and whether the initial observation is preserved.

World Reactivity. The four reaction tasks pair an initial image with one atomic control to probe different forms of world response. **Terrain Interaction** places structured terrain along the subject's path: a horizontal control requires the model to infer height changes and maintain contact on stairs, slopes, or uneven ground. **Object Interaction** places a target where contact is expected, testing displacement of loose objects, deformation of flexible objects, or blockage by rigid objects without interpenetration. **Social Interaction** creates a local conflict with nearby agents and tests their responses, such as yielding, stopping, or changing path. **Physical Reaction** tests how a process unfolds under gravity, friction, momentum transfer, fluid response, or motion constraints. A motion control may trigger the process; under a stop control, an unstable initial scene should still evolve spontaneously while the subject remains stationary. **Goal Completion** requires execution toward a high-level goal, selecting relevant entities and respecting constraints while ignoring distractors.

Camera Control and Scene Revisit form the *static-scene track* across camera-, action-, and language-driven models. The remaining six tasks form the *dynamic-interaction track* for compatible action- and language-driven models; Goal Completion is language-only. Illustrated cases for all eight tasks, with the associated evaluation checklists, appear in Section B.

3.3 Test Case Curation Pipeline

Static-scene cases pair images from existing datasets (Timofte et al., 2017; Ling et al., 2024; Ye et al., 2026) with compositions of atomic camera controls; GPT-5.5 provides scene descriptions. Dynamic-interaction cases require initial scenes tailored to the prescribed controls or goals; we use a generative curation pipeline to construct these test cases (Figure 3).

A task-specific pattern library defines semantic coverage. A schema-guided LLM expands each sampled pattern into a draft containing a scene description, an image-generation prompt, a control

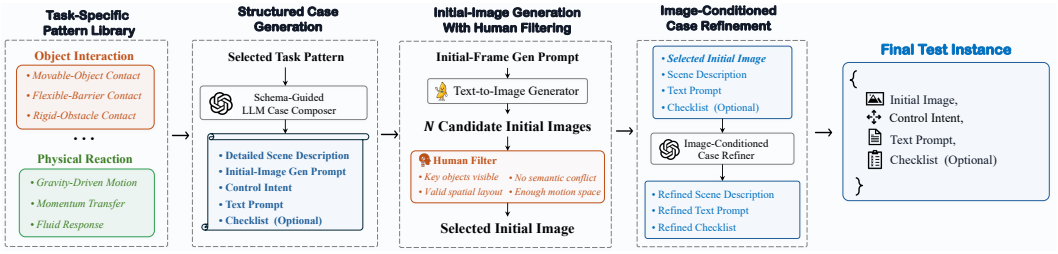


Figure 3: **Test case curation pipeline.** Dynamic-interaction cases are drafted from task patterns, paired with generated and human-filtered initial images, and refined against the selected image.

or goal, and model-facing text. Object Interaction, Social Interaction, Physical Reaction, and Goal Completion additionally include case-specific checklists of observable criteria.

Nano Banana 2 (Team et al., 2023) generates candidate initial images. Human filtering retains visible relevant entities, feasible layouts, and sufficient motion space, rejecting images that already depict the intended event or goal. An image-conditioned refiner revises the scene description, prompts, and checklists to align them with the visible entities and spatial relations in the selected image. Further curation details appear in Section C.1.

Coverage. WorldExam contains 1,474 cases spanning humans, animals, vehicles, and robots across varied scenes and visual styles; Section C.2 details their distributions.

4 Evaluation Protocol and Metrics

WorldExam evaluates Camera Control, Scene Revisit, Subject Control, and Terrain Interaction through reconstructed geometry, and the remaining four tasks through checklist-based judgments. Detailed protocols and metric definitions for static-scene evaluation, displacement calibration, and dynamic-interaction evaluation are provided in Sections D.1 to D.3, respectively.

4.1 Static-Scene Track

Camera Control. Given generated videos, we use VGGT-Ω (Wang et al., 2026a) to estimate camera trajectories. For camera-driven models, the input $\mathrm{SE}(3)$ trajectory serves as the reference. Action- and language-driven inputs specify an ordered sequence of motions without providing an exact 3D camera trajectory. We therefore divide the recovered trajectory into consecutive segments, one for each requested motion, and construct a reference trajectory from the corresponding control. Segment boundaries follow the assigned action durations or, for language inputs, are inferred from changes in the recovered camera motion (Truong et al., 2020). For recovered poses $(\mathbf{R}_t, \mathbf{t}_t)$ and references $(\mathbf{R}_t^*, \mathbf{t}_t^*)$, expressed relative to their starting poses, the errors over T frames are

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

The scale $\hat{s}$ aligns the recovered and reference translations using the interface-specific fitting procedure in Appendix D.1; e_r is measured in degrees. We aggregate these errors over the requested motions and convert them into a Camera Control score:

$$S_{\mathrm{cam}} = 100\sqrt{\max(0, 1 - e_t/0.5)\max(0, 1 - e_r/15)}. \tag{2}$$

Scene Revisit. Scene Revisit checks whether the camera returns to its initial viewpoint and whether the scene remains consistent upon return. For action- and language-driven models, camera motion is not assumed to have constant speed, so equal frame counts for outgoing and return motions do not guarantee a return to the initial viewpoint. To reduce the influence of motion, we search the return segment $\mathcal{T}_{\mathrm{return}}$, selecting the frame whose pose P_t is closest to the initial pose:

$$t_{\mathrm{rev}} = \arg\min_{t \in \mathcal{T}_{\mathrm{return}}} d(P_t, P_1). \tag{3}$$

Here, d is translation distance or relative rotation angle for the corresponding round-trip control. A revisit succeeds within 10% of the maximum outgoing translation or 5° of the initial orientation.

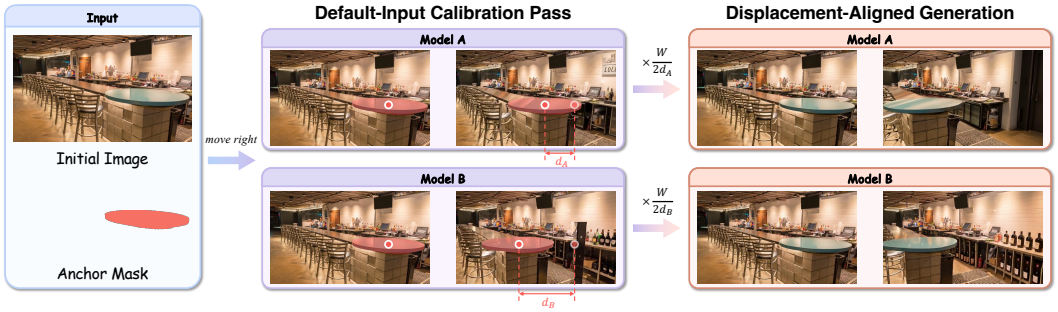


Figure 4: **Image-space displacement alignment.** Calibrating input translations aligns the visible displacement induced by different camera-driven models.

For each case, we compare the selected frame with the input image using PSNR, LPIPS, and SSIM. We normalize these metrics as $s_{\mathrm{P}} = \min(\mathrm{PSNR}/25, 1)$, $s_{\mathrm{L}} = 1 - \mathrm{LPIPS}$, and $s_{\mathrm{S}} = \mathrm{SSIM}$, and combine them with the binary return-success indicator $b \in \{0, 1\}$:

$$S_{\mathrm{rev}} = 100\sqrt{b(s_{\mathrm{P}} + s_{\mathrm{L}} + s_{\mathrm{S}})/3}. \tag{4}$$

The revisit success rate, reported as Revisit Success, is the fraction of cases with $b = 1$.

Image-Space Displacement Alignment for Camera-Driven Models. Identical translation values can induce different image-space displacements across camera-driven models. Larger displacements expose more novel-view content and increase the difficulty of generating controlled, spatially consistent videos. Our ablation in Section 5.4 shows that increasing the input translation leads to declines in Camera Control, Scene Revisit, and general-metric performance. We therefore calibrate input translations to align image-space displacement, as illustrated in Figure 4.

4.2 Dynamic-Interaction Track

Subject Control. Subject Control checks whether the subject follows the requested motion. SAM2 tracks the subject, and VGGT-Ω reconstructs its 3D trajectory, as illustrated in Figure 5. For each control, we compare the corresponding trajectory segment with a horizontal unit direction vector based on the control. We aggregate the errors in Equation (1) to obtain e_t and compute:

$$S_{\mathrm{sub}} = 100 \max(0, 1 - e_t/0.5). \tag{5}$$

Terrain Interaction. Terrain Interaction tests whether horizontal controls induce appropriate vertical adaptation, such as climbing stairs when moving forward. We reconstruct terrain from an image with the subject removed and compare the subject's height changes with those of the terrain beneath its path. For cases that pass the horizontal-control check, we score the percentage of terrain extrema and the endpoint where subject and terrain height changes agree.

Checklist Evaluation. Considering that the remaining four tasks require semantic and causal judgments that cannot be captured by recovered trajectories, we therefore use GPT-5.5 as a VLM judge to check the case-specific checklists constructed in Section 3.3. The VLM judges each item from sampled video frames, checking, for example, whether contact produces an appropriate object response without interpenetration. For video V and checklists $L = \{\ell_k\}_{k=1}^{K}$, the score is the percentage of items supported by clear visual evidence:

$$S_{\mathrm{check}}(V, L) = \frac{100}{K} \sum_{k=1}^{K} \mathbb{I}[\ell_k \text{ is satisfied in } V]. \tag{6}$$

General Metrics and Aggregation. We assess visual quality and temporal consistency with VBench (Huang et al., 2024), adding 3D Consistency (Duan et al., 2025) and Photometric Consistency for static scenes. *Task* and *General* average supported task scores and general metrics, respectively; *Overall* averages the two. Since Goal Completion is language-only, dynamic action and language averages cover five and six tasks, respectively.



Figure 5: **Geometry-based evaluation of Subject Control and Terrain Interaction.** (a) SAM2 and VGGT-Ω recover subject motion relative to estimated gravity. (b) Projecting this trajectory onto reconstructed subject-free terrain provides the corresponding terrain heights.

5 Experiments

We evaluate 23 representative video world models: 6 camera-driven, 10 action-driven, and 7 language-driven models. All 23 enter the static-scene track, and 12 support dynamic-interaction evaluation. Model configurations and task eligibility are detailed in Section E.

5.1 Static-Scene Results

Table 2 reports Camera Control and Scene Revisit alongside the general metrics.

Camera Control. NeoVerse achieves the highest score at 97.33, followed within the camera-driven group by InSpatio-World at 85.94. Both models use explicit 3D guidance through Gaussian and depth priors, suggesting its value for precise camera control. These models, however, are designed to control viewpoints rather than subject–environment interactions, restricting their capacity to the static-scene track. WorldPlay leads the action-driven group at 92.74, with SANA-WM second at 88.74. Language-driven models obtain lower scores, with Hailuo 2.3 reaching 63.29.

Scene Revisit. Reporting return success alongside appearance helps distinguish failure to return from changes in the scene observed upon return. With pose-based search over the return segment, 9 of the 23 models achieve Revisit Success no greater than 0.500, indicating that round-trip control remains challenging. Conventional similarity evaluation at a presumed return frame can therefore conflate viewpoint mismatch with scene changes: low appearance similarity does not necessarily indicate poor spatial consistency. The PSNR, LPIPS, and SSIM values in Table 2 compare the initial frame with the return frame whose recovered pose is closest to the initial pose. Strong Scene Revisit performance therefore requires both reliable returns and faithful appearance recovery. NeoVerse and InSpatio-World combine success rates of 1.000 with the strongest appearance metrics in their group, yielding Scene Revisit scores of 88.97 and 85.63. Among action-driven models, SANA-WM achieves the highest return success (0.935) and combined score (70.73), while WorldPlay leads the appearance metrics. Language-driven models obtain lower Scene Revisit scores, with uniformly low return success (0.325–0.505) identifying round-trip control as a major bottleneck.

5.2 Dynamic-Interaction Results

Table 3 reports Subject Control and the five World Reactivity tasks for compatible models.

Subject Control. LingBot-World-V2 and SANA-WM achieve the highest scores at 80.15 and 77.07, followed by ABot-World at 69.62. All five action-driven models outperform the best language-driven model, Veo 3.1 (37.28), in following the prescribed horizontal subject motion.

Table 2: **Static-scene track evaluation.** Task and General average the two task scores and five general metrics; Overall averages both. Higher is better except e_t, e_r, and LPIPS. Bold/underlined values mark the best/second-best within each interface group.

Model	Camera Control			Scene Revisit					3D Cons.	Photo. Cons.	Temp. Flick.	Aesth. Quality	Imag. Quality	Average		
	$e_t \downarrow$	$e_r \downarrow$	Score↑	Success↑	PSNR↑	LPIPS↓	SSIM↑	Score↑						Task	General	Overall
Camera-driven																
TrajectoryCrafter (Yu et al., 2025)	0.14	1.01	80.32	**1.000**	19.08	0.240	0.545	82.86	96.98	62.83	92.56	52.17	67.03	81.59	74.31	77.95
ReCamMaster (Bai et al., 2025)	0.47	3.91	38.64	0.815	16.02	0.353	0.415	62.14	**99.33**	**81.97**	95.52	54.58	73.45	50.39	**80.97**	65.68
Voyager (Huang et al., 2025a)	0.31	4.29	56.19	0.995	16.88	0.388	0.466	75.75	89.54	27.43	93.04	53.59	63.69	65.97	65.46	65.72
FantasyWorld (Dai et al., 2026)	1.04	7.82	18.46	0.560	15.41	0.323	0.374	42.21	98.57	75.60	**95.91**	**57.11**	**73.98**	30.34	80.23	55.29
NeoVerse (Yang et al., 2026)	**0.01**	**0.60**	**97.33**	**1.000**	**21.72**	**0.141**	**0.662**	**88.97**	98.19	70.83	93.53	52.72	72.19	**93.15**	77.49	**85.32**
InSpatio-World (1.3B) (Team et al., 2026a)	0.09	1.24	85.94	**1.000**	20.79	0.224	0.606	85.63	98.44	64.88	93.68	53.78	73.60	85.79	76.88	81.34
Action-driven																
Hunyuan-GameCraft (Li et al., 2025)	0.09	11.95	41.33	0.385	13.36	0.542	0.367	28.75	93.61	53.93	93.91	54.97	71.69	35.04	73.62	54.33
Astra (Zhu et al., 2026b)	0.31	7.92	32.59	0.365	12.20	0.601	0.309	23.58	96.46	78.41	96.38	51.71	71.78	28.09	78.95	53.52
WorldPlay (Sun et al., 2025b)	**0.04**	**1.04**	**92.74**	0.790	**18.42**	**0.271**	**0.531**	66.99	98.69	81.06	95.90	53.92	73.31	**79.87**	80.58	**80.23**
Yume 1.5 (Mao et al., 2026)	0.12	3.23	75.67	0.255	12.37	0.615	0.359	16.59	98.44	67.53	95.18	53.42	75.28	46.13	77.97	62.05
LingBot-World (Team et al., 2026b)	0.11	6.22	58.19	0.605	15.58	0.341	0.406	45.97	99.59	81.52	**96.53**	**59.87**	75.00	52.08	82.50	67.29
Infinite-World (Wu et al., 2026c)	0.17	2.35	71.70	0.645	14.31	0.388	0.345	46.49	**99.91**	**92.09**	95.99	56.27	**76.41**	59.10	**84.13**	71.62
Matrix-Game 3.0 (Wang et al., 2026b)	0.06	4.36	76.36	0.860	13.82	0.485	0.372	60.26	95.59	30.14	93.40	48.80	73.39	68.31	68.26	68.29
SANA-WM (Zhu et al., 2026a)	**0.04**	1.93	88.74	**0.935**	16.31	0.369	0.447	**70.73**	99.33	75.50	95.36	54.47	73.43	79.74	79.62	79.68
LingBot-World-V2 (Gao et al., 2026)	0.06	7.59	60.69	0.610	12.74	0.516	0.313	41.31	91.79	46.73	92.50	55.12	71.55	51.00	71.54	61.27
ABot-World (Jiang et al., 2026)	0.12	5.39	64.81	0.845	11.94	0.664	0.372	53.30	82.79	10.30	95.38	47.31	65.64	59.06	60.28	59.67
Language-driven																
Kling 2.5 (Kuaishou Technology, 2025)	0.19	6.90	50.18	0.325	**13.54**	0.532	0.381	23.50	**99.87**	**87.87**	**97.56**	56.28	75.70	36.84	**83.46**	60.15
Veo 3.1 (Google DeepMind, 2025)	0.27	8.22	40.83	0.447	12.30	0.594	0.345	28.77	99.06	73.61	95.43	55.49	76.57	34.80	80.03	57.42
Hailuo 2.3 (MiniMax, 2024)	**0.15**	**5.20**	**63.29**	**0.505**	13.52	**0.519**	**0.387**	**36.45**	99.38	79.88	95.33	**56.30**	75.52	**49.87**	81.28	**65.58**
Wan 2.6 I2V (Wan et al., 2025)	0.17	6.12	57.72	0.495	13.28	0.527	0.353	35.20	99.49	69.42	94.13	53.93	**77.53**	46.46	78.90	62.68
Seedance 1.5 (Seedance et al., 2025)	0.20	7.69	49.18	0.375	12.28	0.602	0.342	23.69	97.31	56.60	94.81	54.45	74.88	36.44	75.61	56.03
Vidu Q3 (Bao et al., 2024)	0.24	8.47	42.75	0.360	12.65	0.586	0.351	24.51	99.43	71.28	95.01	55.38	76.73	33.63	79.57	56.60
HappyHorse 1.0 (Alibaba Group, 2026)	0.18	5.62	58.29	0.420	12.48	0.564	0.352	28.82	99.62	74.17	95.02	55.56	77.00	43.56	80.27	61.92

Table 3: **Dynamic-interaction track evaluation.** Task and General average supported task scores and four general metrics; Overall averages both. Higher is better. Dashes denote unsupported tasks. Bold/underlined values mark the best/second-best within each interface group.

Model	Subject Control	Terrain Inter.	Object Inter.	Social Inter.	Phys. React.	Goal Comp.	Subject Cons.	Motion Smooth.	Aesth. Quality	Imag. Quality	Average		
											Task	General	Overall
Action-driven													
WorldPlay (Sun et al., 2025b)	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 (Team et al., 2026b)	55.47	24.33	25.94	**60.37**	33.43	–	**94.98**	**98.89**	60.86	**71.69**	39.91	**81.61**	60.76
SANA-WM (Zhu et al., 2026a)	77.07	27.28	**40.19**	48.71	36.41	–	92.37	98.55	55.79	71.28	45.93	79.50	62.71
LingBot-World-V2 (Gao et al., 2026)	**80.15**	44.63	31.24	45.71	32.83	–	89.27	97.99	**61.39**	71.36	46.91	80.00	**63.46**
ABot-World (Jiang et al., 2026)	69.62	**46.74**	36.88	48.08	**39.17**	–	87.59	98.26	57.54	67.72	**48.10**	77.78	62.94
Language-driven													
Kling 2.5 (Kuaishou Technology, 2025)	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 (Google DeepMind, 2025)	**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 (MiniMax, 2024)	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 (Wan et al., 2025)	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 (Seedance et al., 2025)	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 (Bao et al., 2024)	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 (Alibaba Group, 2026)	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

Terrain Interaction. The horizontal input control leaves vertical motion unspecified, requiring the model to infer how the subject should adapt to the terrain. Vidu Q3 achieves the highest score at 64.39, followed by Hailuo 2.3 at 61.57; ABot-World leads the action-driven group at 46.74. The different leaders in Subject Control and Terrain Interaction indicate that many action-driven models follow explicit instructions well, but struggle in inferring how to adapt their motion to the terrain.

Object Interaction, Social Interaction, and Physical Reaction. With expected consequences withheld from the inputs, these scores assess scene-grounded reactions. Veo 3.1 leads Object and Social Interaction at 75.96 and 85.10, whereas Hailuo 2.3 leads Physical Reaction at 63.84. Within the action-driven group, the respective leaders are SANA-WM, LingBot-World, and ABot-World. Case-level checklists expose failures: a cart penetrates a rigid wall, and crowd interactions break down. Physical Reaction also reveals missing scene-implied dynamics, such as an unstable basket remaining supported instead of falling; paired sequences and checklist judgments appear in Section B.

Goal Completion. Goal Completion extends the evaluation to execution toward a stated target, with intermediate steps left unspecified. HappyHorse 1.0 and Veo 3.1 achieve the highest checklist scores at 85.33 and 85.30. Case-level checklists also distinguish performing a relevant action from achieving its intended outcome, such as grasping a distractor instead of the requested object; Section B.8 contrasts successful placement with incorrect object selection and incomplete execution.

Table 4: **Human alignment.**

Task	Checklist Items	Spearman ρ	PLCC
Goal Comp.	1,303	0.8960	0.9017
Phys. React.	1,425	0.8838	0.8750
Object Inter.	1,527	0.8251	0.8541
Social Inter.	1,538	0.7019	0.7103
Overall	**5,793**	**0.8614**	**0.8583**

Table 5: **Input translation multiplier ablation.**

Translation Multiplier	Camera Control	Scene Revisit	3D Cons.	Photo. Cons.	Temp. Flick.	Aesth. Quality	Imag. Quality	Average		
								Task	General	Overall
0.10×	98.25	90.60	99.91	80.17	95.41	54.06	72.53	94.43	80.42	87.43
0.50×	97.84	89.81	99.54	76.33	94.21	53.39	72.47	93.83	79.19	86.51
0.75×	97.57	89.50	99.07	73.86	93.81	53.03	72.30	93.54	78.41	85.98
1.00×	97.33	88.97	98.19	70.83	93.53	52.72	72.19	93.15	77.49	85.32
2.00×	95.32	88.06	96.32	62.45	92.93	52.00	71.54	91.69	75.05	83.37

5.3 Cross-Task Analysis

Control Adherence and Scene-Conditioned Reactions. Unlike Subject Control, the cases for four reaction tasks are only paired with one single atomic control for less control complexity and more focused evaluation of scene-conditioned reactions. All action-driven models outperform language-driven models on Subject Control, but trail on Social Interaction (Table 3 and Figure 16). Moreover, Hailuo 2.3 and Vidu Q3 exceed all action-driven models on Terrain Interaction, Object Interaction, and Physical Reaction. These results show that strong control adherence does not consistently extend to scene-conditioned reactions, while stronger reactions can coexist with less reliable control. Further progress requires combining these strengths, so that models both execute the requested action and generate its unstated consequences for the surrounding world.

Visual Quality and Task Performance. General visual metrics do not fully capture these behavioral differences: language-driven models have similar General averages (79.64–81.04), yet their Task averages vary substantially (39.85–65.02). This gap motivates assessing visual quality and interaction correctness separately: convincing appearance does not ensure appropriate reactions.

5.4 Evaluator Validation and Sensitivity

Human Alignment. To validate the checklist evaluator against human judgments, we sample 200 generated videos per checklist-based task from the outputs of multiple models. Three annotators independently assess each video using the same 10 frames as GPT-5.5. Human scores are the per-video fractions of checklist items satisfied by majority vote. Table 4 shows strong agreement between automated and human scores, with pooled Spearman $\rho = 0.8614$ and PLCC $= 0.8583$.

Input Translation Multiplier. We evaluate NeoVerse under the same camera controls while varying only the translation multiplier of its input $\mathrm{SE}(3)$ trajectory. As the multiplier increases from $0.10\times$ to $2.00\times$, Camera Control decreases from 98.25 to 95.32, Scene Revisit from 90.60 to 88.06, and Photometric Consistency from 80.17 to 62.45 (Table 5). Revisit Success remains 1.000 at all five multipliers, so the Scene Revisit decline reflects appearance degradation despite successful returns. 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 Section 4.1, which is applied to all reported camera-driven results.

Reconstruction Backend. Replacing VGGT-Ω with DA3 (Lin et al., 2025) preserves the broad performance patterns: strong-performing models remain competitive, and the distinction between control adherence and scene-conditioned reactions persists. The main conclusions of our cross-task analysis therefore remain unchanged across the two backends; full results appear in Section F.3.

6 Conclusion

We introduced WorldExam, a unified benchmark across three interfaces that pairs foundational capability tests with systematic diagnosis of terrain, object, and social interactions, physical reactions, and goal completion. Across 23 models, strong control adherence does not guarantee appropriate scene-conditioned reactions, while similar visual quality conceals substantial differences in task performance. Human agreement and consistent cross-task analysis across reconstruction backends support evaluation reliability. These findings suggest that future world-model development should give equal attention to control adherence and inherent reactivity, so that generated worlds both execute instructions and produce coherent scene-conditioned consequences.

AI USE STATEMENT

In this work, we use generative AI for synthetic case construction, scene annotation, and checklist-based evaluation. Dynamic-interaction cases require initial scenes in which prescribed controls can trigger the intended interactions or support a specified goal. We use LLMs to draft cases and Nano Banana 2 to generate initial images; GPT-5.5 describes scenes, and image-conditioned refinement grounds prompts and checklists in the selected images (Section 3.3). Considering that Object Interaction, Social Interaction, Physical Reaction, and Goal Completion require semantic and causal judgments that cannot be captured by recovered trajectories, we use GPT-5.5 as a VLM judge to check the case-specific checklists, enabling evaluation at benchmark scale (Section 4.2). We also adopt human filtering to check candidate images for relevant entities, feasible layouts, and the required initial state, and three human annotators assess sampled outputs to validate checklist scoring (Section 5.4). Additionally, we use generative AI to polish the manuscript for clarity and readability. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The static-scene track uses scene images spanning diverse visual styles collected from Flickr2K, DL3DV-10K, and MIND (Timofte et al., 2017; Ling et al., 2024; Ye et al., 2026). The selected images contain no faces or other privacy-sensitive information. We explicitly acknowledge the source datasets, and third-party images remain subject to their original licenses and terms of use.

Three members of the author team performed image annotation and validation, including checking candidate images for relevant entities, feasible layouts, and the required initial state. The same three authors independently assessed sampled model outputs against case-specific checklists to validate automated scoring. These activities involved a manageable workload within the author team.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe benchmark construction, interface adaptation, and evaluation metrics. Appendices C, D, and E provide curation details and benchmark statistics, detailed scoring protocols, and per-model generation settings. Paired examples and checklists appear in Appendix B, and additional experimental results in Appendix F. The supplementary material includes five test cases per task (40 in total) as a reference subset illustrating the data format and evaluation inputs. We plan to release the evaluation code and the full benchmark, including all 1,474 cases, upon publication.

REFERENCES

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

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, volume 2025, pp. 102075–102121, 2025.

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.

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. https://openai.com/index/

video-generation-models-as-world-simulators/, 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/.

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.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *International Conference on Learning Representations*, volume 2026, pp. 103603–103622, 2026.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 27713–27724, 2025.

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.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

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

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 Transactions on Graphics (TOG)*, 44(6): 1–15, 2025a.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. Abot-world-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026.

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

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36385–36399, 2026.

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.

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.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

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.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026.

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.

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

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 *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

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.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416. IEEE, 2025a.

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*, 2025b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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*, 2026a.

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*, 2026b.

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 *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1110–1121. IEEE, 2017.

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

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.

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.

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*, 2026a.

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*, 2026b.

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*, 2026a.

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*, 2026b.

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*, 2026c.

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*, 2026a.

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

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.

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.

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.

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.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025.

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.

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.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

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 *International Conference on Learning Representations*, volume 2026, pp. 79167–79184, 2026b.

SUPPLEMENTARY MATERIAL

The appendix presents representative scenes and paired task examples, followed by dataset specifications, evaluation details, model settings, and additional results.

CONTENTS

A Gallery 16

B Qualitative Examples of the Eight Tasks 17

- B.1 Camera Control 17
- B.2 Subject Control 17
- B.3 Scene Revisit 18
- B.4 Terrain Interaction 18
- B.5 Object Interaction 19
- B.6 Social Interaction 20
- B.7 Physical Reaction 21
- B.8 Goal Completion 22

C Benchmark Curation and Statistics 23

- C.1 Detailed Curation Pipeline 23
- C.2 Benchmark Statistics 23

D Detailed Evaluation Protocol 24

- D.1 Static-Scene Track 24
- D.2 Image-Space Displacement Alignment 25
- D.3 Dynamic-Interaction Track 26

E Per-Model Inference Settings 27

F Additional Experimental Results 28

- F.1 Cross-Task Profiles 28
- F.2 Input Translation Multiplier Ablation 28
- F.3 Sensitivity to the Reconstruction Backend 29

A GALLERY

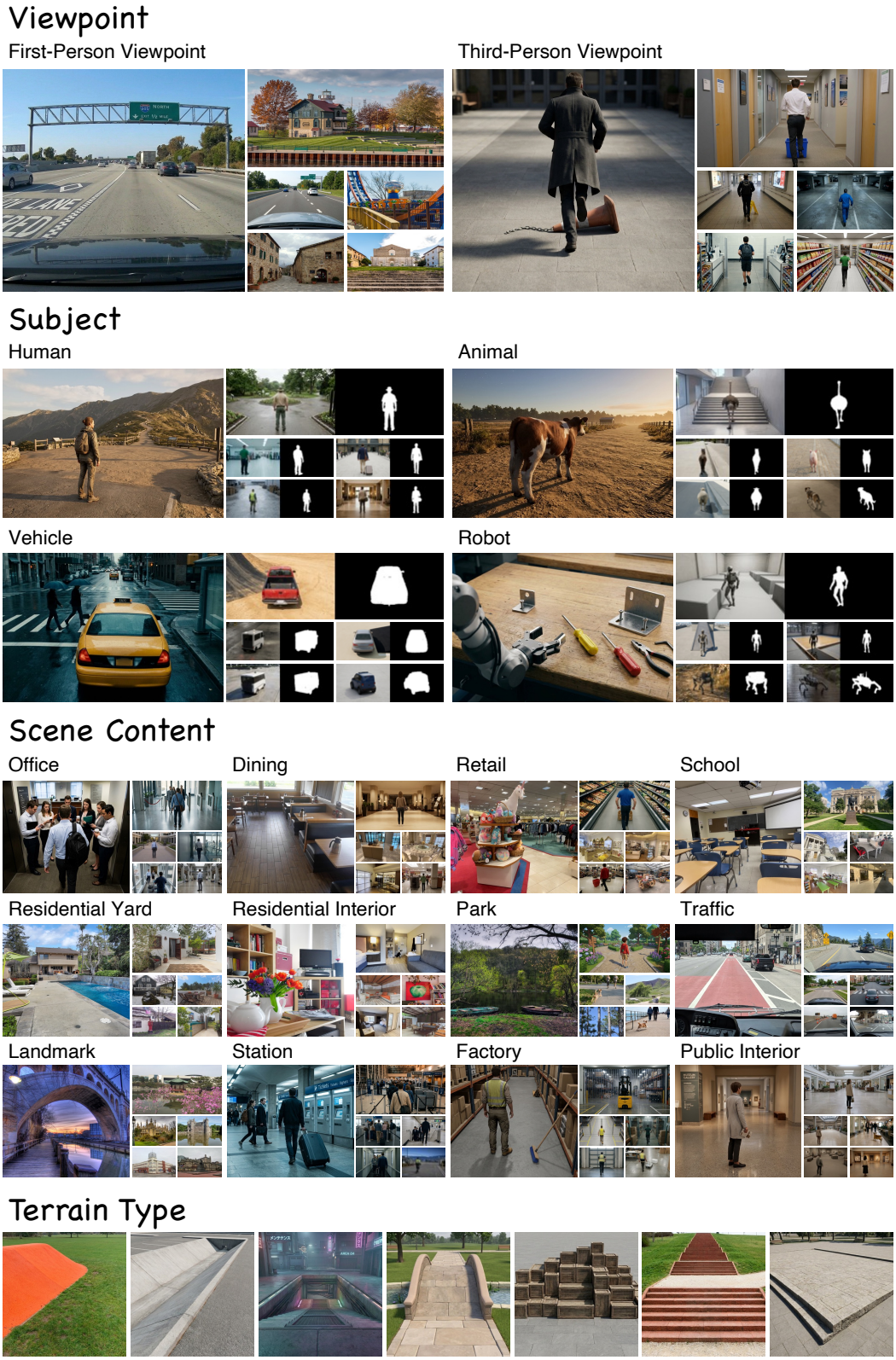


Figure 6: **WorldExam Gallery.** Cases span diverse viewpoints, subjects, scenes, and terrain.

B Qualitative Examples of the Eight Tasks

Each pair shares the initial scene and control intent. Frames are sampled nearest the labeled fractions of the final frame time T. **Full videos appear in the supplementary viewer**.

B.1 Camera Control

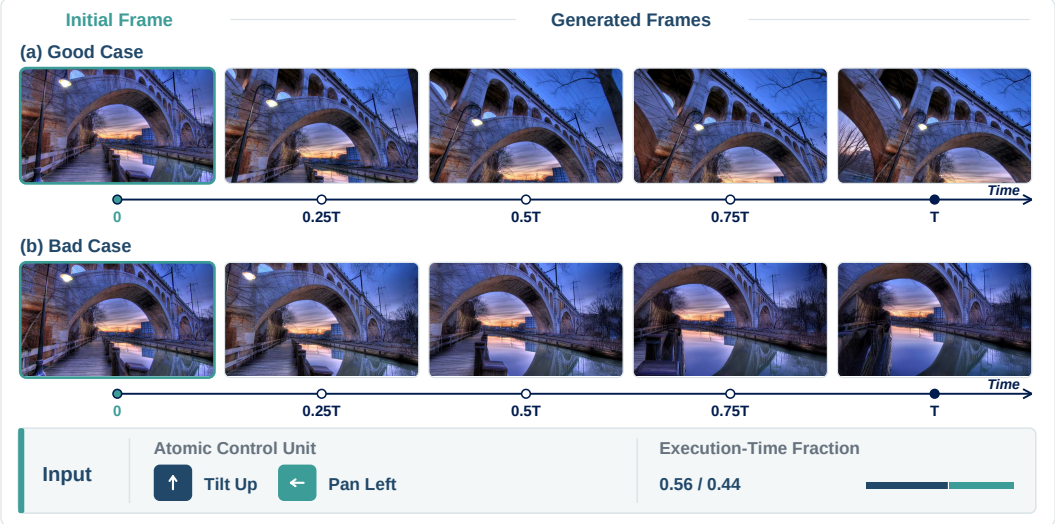


Figure 7: **Camera Control.** Paired outputs.

The good case follows the upward tilt and subsequent leftward pan, while the bad case drifts toward the right without completing the prescribed rotations.

B.2 Subject Control

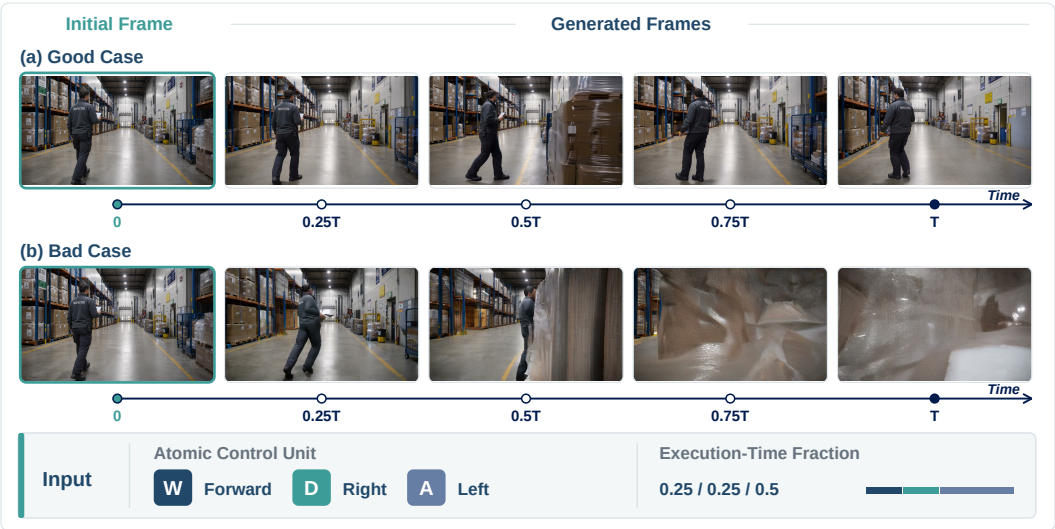


Figure 8: **Subject Control.** Paired outputs.

The good case preserves the subject through all three motion segments. In the bad case, the subject disappears as the view enters the boxes; tracking is lost in 56 of 125 frames, exceeding the validity threshold and yielding a zero score.

B.3 SCENE REVISIT

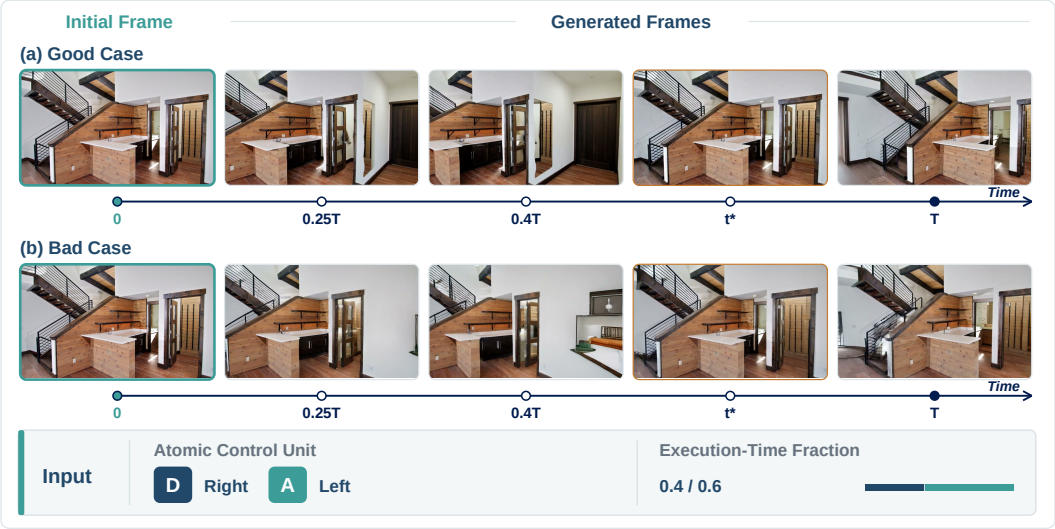


Figure 9: **Scene Revisit.** t^* denotes the pose-selected return frame in each video.

Both outputs pass the geometric return criterion. At the pose-selected frames, the good case recovers the entryway more faithfully, while the bad case alters the staircase, cabinetry, and doorway.

Case	Frame (0-based)	Success	PSNR	LPIPS	SSIM
Good Case	64	Yes	20.07	0.081	0.770
Bad Case	59	Yes	12.64	0.364	0.392

B.4 TERRAIN INTERACTION

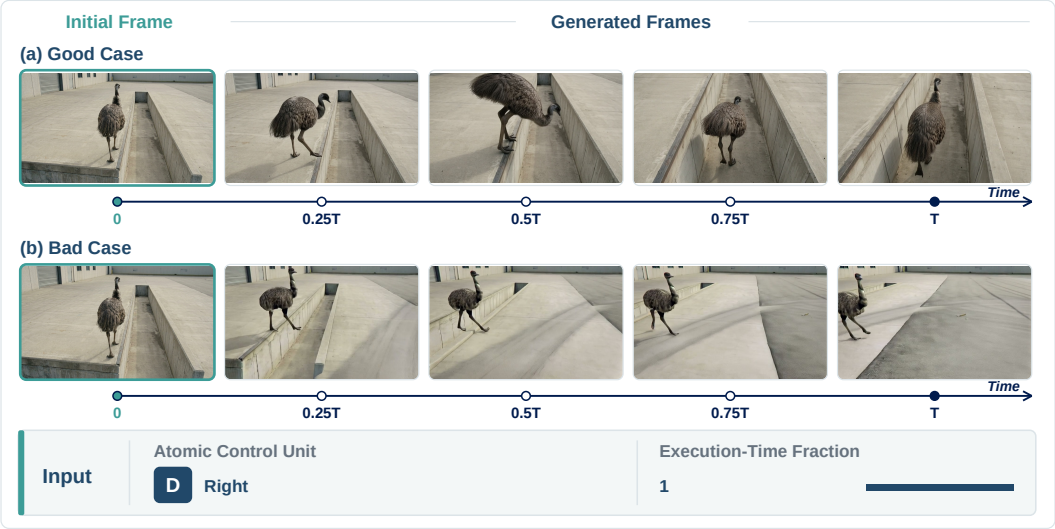


Figure 10: **Terrain Interaction.** Paired outputs.

The good case descends into the trench and matches terrain height; the bad case shows insufficient descent.

B.5 Object Interaction

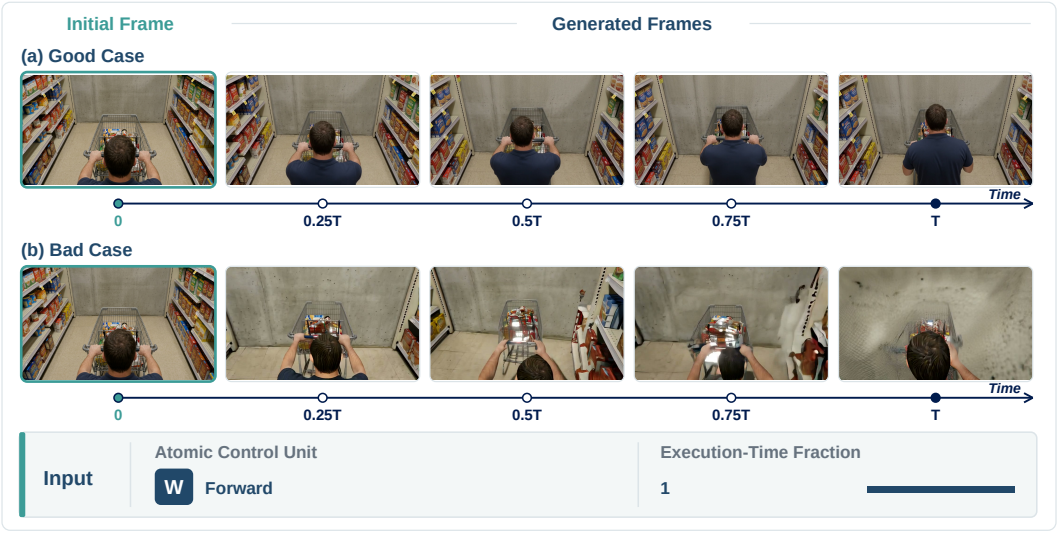


Figure 11: **Object Interaction.** Paired outputs.

The good case stops the cart at the wall. The bad case reaches the obstacle but continues into a deforming wall.

Case-Specific Evaluation Checklist

Evaluation only; withheld from model inputs. Columns (a) and (b) match the rows above; ✓ / ✗ indicate satisfied / unsatisfied items.

#	Case-specific criterion	(a)	(b)
1	Does the shopper continue walking straight forward behind the cart on the same aisle centerline until the cart reaches the wall?	✓	✓
2	Does the cart's front rail or front corners make direct contact with the concrete wall rather than missing it or striking a shelf first?	✓	✓
3	After contact, does the wall stay fixed in place with no visible shift, swing, or sliding response?	✓	✗
4	After the cart meets the wall, is the cart's forward travel visibly stopped or sharply arrested by the barrier?	✓	✗
5	Does the shopper's upper body carry forward into the handle area or show a brief forward compression when the cart is blocked?	✓	✗
6	Does the cart remain generally square to the wall at impact instead of glancing off and rotating away immediately?	✓	✓
7	Does the continuation avoid any larger competing collision event, keeping the cart-to-wall stop as the main physical interaction?	✓	✗
8	Does the camera keep the cart front and wall contact zone visible enough during and just after impact to judge the stoppage clearly?	✓	✓
	Satisfied items	8/8	4/8

B.6 SOCIAL INTERACTION



Figure 12: **Social Interaction.** Paired outputs.

The good case preserves a coherent crowd with small head and body turns. The bad case also shows early turns, satisfying the response item, but later loses bystander visibility and introduces body distortions and clipping.

Case-Specific Evaluation Checklist

Evaluation only; withheld from model inputs. Columns (a) and (b) match the rows above; ✓ / ✗ indicate satisfied / unsatisfied items.

#	Case-specific criterion	(a)	(b)
1	Does the pedestrian continue into the packed awning shelter space?	✓	✓
2	Do the six sheltering pedestrians remain visible enough to judge their reactions?	✓	✗
3	Does at least one nearby pedestrian visibly squeeze over, turn, or make space under the awning?	✓	✓
4	Does the interaction remain centered on the storefront shelter zone?	✓	✓
5	Is the protagonist's intrusion into the compact waiting area clearly sustained?	✓	✓
6	Do the reactions fit a realistic rainy storefront crowd?	✓	✗
7	Does the continuation stay socially coherent rather than treating the shelter as open empty space?	✓	✗
	Satisfied items	7/7	4/7

B.7 PHYSICAL REACTION



Figure 13: **Physical Reaction.** Paired outputs.

The good case continues the basket's outward rotation and fall while the woman stays in place without touching it. The bad case instead returns the basket toward a level, supported position, following the no-motion control without generating the expected physical evolution.

Case-Specific Evaluation Checklist

Evaluation only; withheld from model inputs. Columns (a) and (b) match the rows above; ✓ / ✗ indicate satisfied / unsatisfied items.

#	Case-specific criterion	(a)	(b)
1	Does the woman remain stationary throughout the short continuation?	✓	✓
2	Does the basket continue rotating outward and downward from its already tilted position?	✓	✗
3	Does the basket fall off the washing machine rather than sliding back to a fully supported position on top?	✓	✗
4	Does the motion begin as a tip about the washer edge or remaining contact point, consistent with overhanging weight pulling it over?	✓	✗
5	Does the overhanging towel load move with the basket and contribute to the same outward fall direction?	✓	✗
6	Does the basket fall without any new touch, grab, or bump from the woman?	✓	✗
7	After leaving support, does the basket and laundry move downward into the open space in front of the machine rather than floating or reversing upward?	✓	✗
	Satisfied items	7/7	1/7

B.8 GOAL COMPLETION

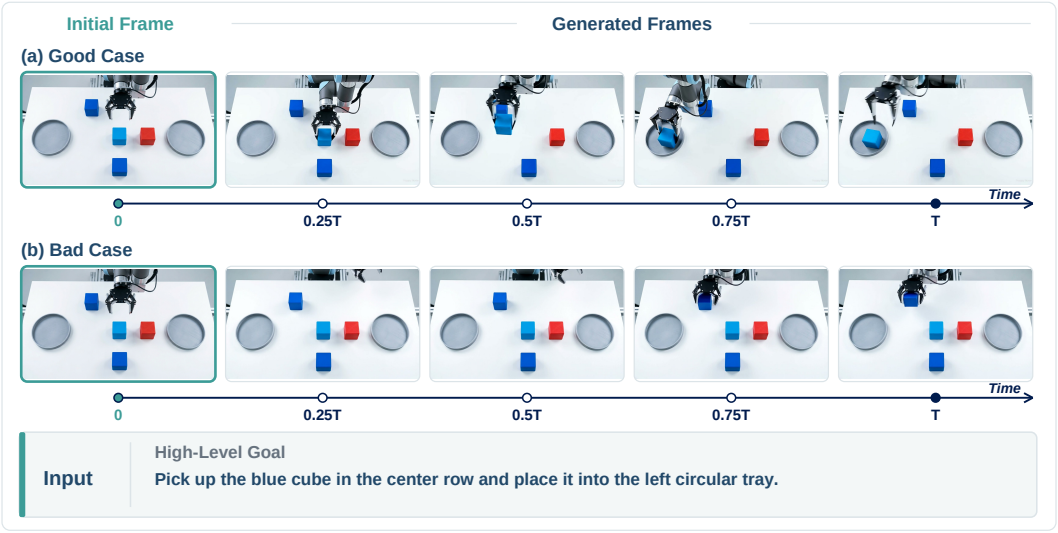


Figure 14: **Goal Completion.** Paired outputs.

The good case places the light-blue target in the left tray. The bad case grasps the upper-left dark-blue cube and ends without placement, failing the target-selection and completion criteria.

Case-Specific Evaluation Checklist

Evaluation only; withheld from model inputs. Columns (a) and (b) match the rows above; ✓ / ✗ indicate satisfied / unsatisfied items.

#	Case-specific criterion	(a)	(b)
1	Does the robot grasp the light-blue cube in the center row?	✓	✗
2	Does the robot avoid grasping either of the other blue cubes?	✓	✗
3	Does the robot lift the target cube clear of the table before transporting it?	✓	✗
4	Does the robot move the held cube toward the left circular tray?	✓	✗
5	Does the robot avoid placing the cube into the right circular tray?	✓	✓
6	Does the robot release the cube inside the left circular tray?	✓	✗
7	Is the light-blue center-row cube visibly resting within the left tray at the end?	✓	✗
	Satisfied items	7/7	1/7

C Benchmark Curation and Statistics

C.1 Detailed Curation Pipeline

This section expands the construction process in Section 3.3 and Figure 3. For the static-scene track, we select scene images from Flickr2K, DL3DV-10K, and MIND (Timofte et al., 2017; Ling et al., 2024; Ye et al., 2026), obtain scene descriptions with GPT-5.5, and pair the images with atomic camera controls. The six dynamic-interaction tasks instead use the following construction process to obtain initial scenes that support the intended controls or goals.

Pattern-Guided Drafting. A task-specific pattern library covers subject motion, structured terrain, object contact, social conflict, physical processes, and goal-directed situations. For each sampled pattern, a schema-guided LLM composer produces a structured draft containing a scene description, an initial-image generation prompt, a control intent or high-level goal, and a draft text prompt for language-driven models. Object Interaction, Social Interaction, Physical Reaction, and Goal Completion also include case-specific checklists of observable criteria. The model-facing text specifies only the control or goal; reactions and execution steps remain unstated.

Candidate Generation and Human Filtering. Nano Banana 2 (Team et al., 2023) generates candidate initial images from the image-generation prompt. Human reviewers retain candidates whose relevant entities are visible, whose layouts support the intended behavior, and whose remaining space permits the subsequent motion. They discard candidates that hide required entities, contradict the draft, make the behavior physically infeasible, or already depict the evaluated event or completed goal. The selected image must provide a pre-event state that supports the intended interaction or goal.

Image-Conditioned Refinement. A case refiner revises the draft against the selected image while preserving the sampled pattern and intended control or goal. It updates entity references, spatial relations, scene descriptions, model-facing text, and optional checklists so that the required entities and preconditions are grounded in the actual image. This avoids criteria based only on the draft. For checklist-based tasks, criteria are fixed before generated videos are evaluated.

Finalized Cases and Model Inputs. Each finalized case contains an initial image, a control intent or high-level goal, a grounded text prompt, and an optional checklist. The control is adapted to the evaluated model's interface as described in Section 3. The image-generation prompt is used only for constructing the initial image, and the checklist is reserved for evaluation; neither is supplied to the evaluated video model. Illustrated cases and their associated criteria appear in Section B.

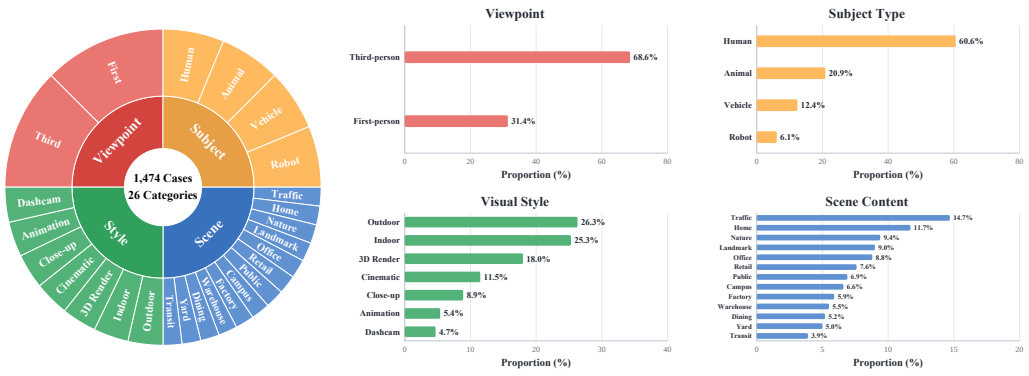
Auxiliary Annotations. Human annotators use SAM-assisted annotation to delineate scene anchors for input-displacement calibration and subject masks for subsequent tracking. They also manually box horizontal surface regions to create horizontal-region masks for estimating the scene's horizontal plane. For Terrain Interaction, we provide Nano Banana 2 with the initial image containing the subject and ask it to remove the subject, producing a subject-free image for terrain reconstruction. These auxiliary assets support the calibration and geometry-based evaluation in Section D.

C.2 Benchmark Statistics

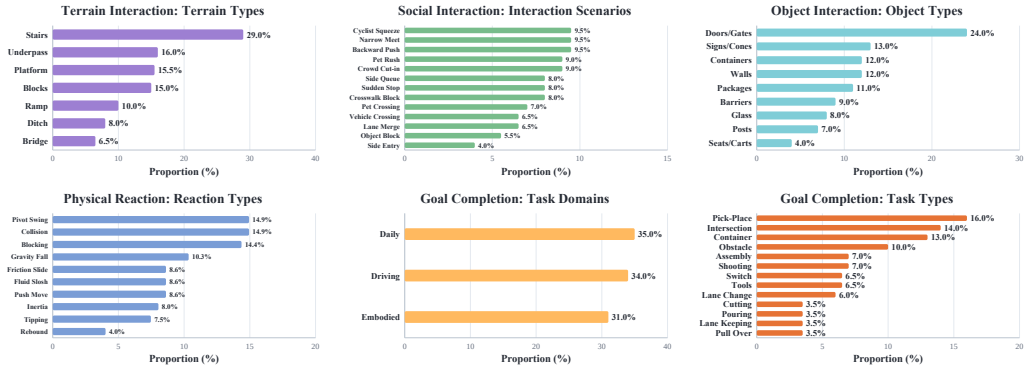
Figure 15 summarizes both the overall dataset composition and the task-specific composition of the five *World Reactivity* tasks.

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 Section A.



(a) Dataset composition.



(b) World Reactivity composition.

Figure 15: **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.

D Detailed Evaluation Protocol

This section supplies the implementation details for the metrics in Section 4; task examples and case-specific checklists appear in Section B.

D.1 Static-Scene Track

Camera Control. Each case specifies an ordered sequence $\{(a_i, \rho_i)\}_{i=1}^{N} (1 \le N \le 3)$. Here, $a_i \in \{\text{W, S, A, D, }\uparrow, \downarrow, \leftarrow, \rightarrow\}$ 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 $\{(\mathbf{R}_t, \mathbf{t}_t)\}_{t=1}^{T}$. Since 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 $\mathrm{SE}(3)$. 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. For this interface, we fit the translation scale by minimizing the mean Euclidean distance:

$$\hat{s} = \arg\min_{s \ge 0} \frac{1}{T} \sum_{t=1}^{T} \| s\mathbf{t}_t - \mathbf{t}_t^* \|_2. \tag{7}$$

We then compute the translation and rotation errors using Equation (1); 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]." The prompt specifies control order but leaves durations open.

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 (Truong et al., 2020) 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 these action- and language-driven translation segments, we first fit the scale by least squares and then compute the mean Euclidean error:

$$\hat{s} = \arg\min_{s \geq 0} \sum_{t=1}^{T} \|s\mathbf{t}_t - \mathbf{t}_t^*\|_2^2, \qquad e_t = \frac{1}{T} \sum_{t=1}^{T} \|\hat{s}\mathbf{t}_t - \mathbf{t}_t^*\|_2. \tag{8}$$

Here, T denotes the segment length. The squared residuals are used only to fit $\hat{s}$; the reported e_t is neither a mean squared error nor an RMSE. For a rotation control, we set $\mathbf{t}^* = \mathbf{0}$ and construct $\mathbf{R}^*$ using the prescribed axis and direction. Its translation error is the mean per-frame $\|\mathbf{t}_t\|_2$ relative to the segment's initial pose, without scale fitting (equivalently, $\hat{s} = 1$ in Equation (1)). 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.

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. We average segment errors with weights proportional to their frame counts to obtain the video-level errors (e_t, e_r).

Scene Revisit. 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. We select the closest pose within this window using Equation (3), apply the return criterion in Section 4.1.

General Metrics. Across all static-scene videos, we additionally report five task-agnostic general metrics. 3D Consistency adapts the metric of WorldScore (Duan et al., 2025) 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 (Huang et al., 2024).

D.2 Image-Space Displacement Alignment for Camera-Driven Models

For each model m and case c, we use an anchor mask provided on the initial image to estimate a translation calibration factor. We first generate a calibration video using the model's default input translation magnitude and a horizontal camera control, either "move left" or "move right."

SAM2 (Ravi et al., 2025) tracks the anchor through the video. Let $d_{m,c}$ denote the horizontal image-space displacement of the tracked mask center and W the frame width. We set the target displacement to half the image width and compute the calibration factor as

$$d^* = W/2, \qquad k_{m,c} = \frac{d^*}{d_{m,c}} = \frac{W}{2d_{m,c}}. \tag{9}$$

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.

D.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 Figure 5, SAM2 (Ravi et al., 2025) tracks the designated subject from a first-frame mask, and VGGT-Ω (Wang et al., 2026a) lifts the tracked pixels into 3D to recover the subject trajectory. The horizontal-region mask estimates gravity and the horizontal plane. The forward direction is the initial camera's viewing direction projected onto this plane; backward reverses it, and left and right are perpendicular directions within this estimated horizontal plane.

Before segment scoring, a case receives zero if the subject mask is missing or empty in more than 33.33% of tracked frames, if its reconstructed trajectory or scene scale is invalid, or if its total trajectory length is below 0.5% of the scene scale. This whole-video check uses cumulative path length, so a subject that moves and returns to its starting position is not treated as stationary.

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 use the least-squares scale fit and mean Euclidean error in Equation (8) to obtain 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 $e_{t,i} = 0.5$. Frame-count-weighted averaging yields e_t for the Subject Control score in Equation (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. Projecting the subject trajectory along gravity gives the corresponding terrain-height profile.

We first require a horizontal Subject Control score of at least 10 (on the 0–100 scale), so that height adaptation is evaluated only when the prescribed horizontal motion is sufficiently executed. Cases with invalid tracking or geometry, or fewer than three valid terrain-projected points, receive zero. For the remaining cases, we median-smooth the subject and terrain height traces with a nine-frame window. We then sample the terrain trace at up to 20 evenly spaced time indices and select its significant extrema and endpoint for evaluation. At the selected terrain-height extrema and endpoint, we assess whether the subject's height changes follow the same trend as the terrain. The case score is the percentage of selected points satisfying this requirement.

Checklist Judgments. GPT-5.5 receives 10 temporally ordered frames uniformly sampled from each generated video and the fixed case-specific checklist. Each item is marked satisfied only when supported by clear visual evidence in the sampled frames; otherwise, it is marked unsatisfied. The checklist score in Section 4.2 averages these decisions within each case, crediting observable partial success.

For *Object Interaction*, criteria verify that contact with the designated object causes a type-consistent reaction whose direction and extent match the contact, without interpenetration. For *Social Interac-*

tion, they check whether the controlled motion induces a timely, attributable adjustment by a visible affected agent, rather than an unchanged, delayed, or unrelated response. For *Physical Reaction*, they assess causal timing and evolution under the relevant physical regularity while preserving required supports, attachments, contacts, and constraints. For *Goal Completion*, they assess correct grounding, intermediate progress, compliance with ordering and scene constraints, and final completion; alternative executions receive credit when they satisfy the same observable requirements. Paired examples and their item-level judgments appear in Sections B.5 to B.8.

General Metrics. The dynamic-interaction track separately reports four VBench metrics (Huang et al., 2024): 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.

E 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. Table 6 lists the evaluated checkpoints, resolutions, frame counts, and task eligibility. Eligibility for the dynamic-interaction track requires reliable control of a visible third-person subject. Models that do not support reliable third-person subject control are restricted to the static-scene track; Goal Completion is evaluated only through language inputs.

Table 6: **Per-model inference settings and eligibility for the dynamic-interaction track.** *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 (Yu et al., 2025)	Local	672×384	49	–
ReCamMaster (Bai et al., 2025)	Local	832×480	81	–
Voyager (Huang et al., 2025a)	Local	768×512	49	–
FantasyWorld (Dai et al., 2026)	Local	592×336	81	–
NeoVerse (Yang et al., 2026)	Local	560×336	81	–
InSpatio (1.3B) (Team et al., 2026a)	Local	832×480	81	–
Action-driven				
Hunyuan-GameCraft (Li et al., 2025)	Local	1216×704	132	✗ TPV†
Astra (Zhu et al., 2026b)	Local	832×480	161	✗ FPV only
WorldPlay (Sun et al., 2025b)	Local	832×480	125	✓ TPV
Yume 1.5 (Mao et al., 2026)	Local	1280×704	145	✗ FPV only
LingBot-World (Team et al., 2026b)	Local	832×464	161	✓ TPV
Infinite-World (Wu et al., 2026c)	Local	896×448	161	✗ FPV only
Matrix-Game 3.0 (Wang et al., 2026b)	Local	1280×704	177	✗ FPV only
SANA-WM (Zhu et al., 2026a)	Local	1280×704	161	✓ TPV
LingBot-World-V2 (Gao et al., 2026)	Local	832×464	157	✓ TPV
ABot-World (Jiang et al., 2026)	Local	1280×704	141	✓ TPV
Language-driven				
Kling 2.5 (Kuaishou Technology, 2025)	API	1280×720	153	✓ TPV prompt
Veo 3.1 (Google DeepMind, 2025)	API	1280×720	192	✓ TPV prompt
Hailuo 2.3 (MiniMax, 2024)	API	1024×768	141	✓ TPV prompt
Wan 2.6 I2V (Wan et al., 2025)	API	1280×720	150	✓ TPV prompt
Seedance 1.5 (Seedance et al., 2025)	API	1280×720	97	✓ TPV prompt
Vidu Q3 (Bao et al., 2024)	API	1280×720	121	✓ TPV prompt
HappyHorse 1.0 (Alibaba Group, 2026)	API	1280×720	123	✓ TPV prompt

F Additional Experimental Results

F.1 Cross-Task Profiles

Figure 16 visualizes two models per interface from Tables 2 and 3, extending the cross-task analysis in Section 5.3. The profiles show complementary strengths: camera-driven models provide precise viewpoint control and scene revisiting, action-driven models offer stronger subject control, and language-driven models show stronger scene-conditioned reactions. Across the full dynamic-interaction results, all five action-driven models exceed all seven language-driven models on Subject Control, whereas all seven language-driven models exceed the action-driven models on Social Interaction. Since the reaction tasks use a single atomic control and leave the expected consequences unstated, this contrast highlights the distinction between executing a prescribed motion and generating the surrounding world's response.

General visual metrics capture a different aspect of these capabilities. Among language-driven models, General averages span only 79.64–81.04, while Task averages range from 39.85 to 65.02. Thus, similarly high visual quality can accompany substantially different control and interaction performance, supporting separate assessment across the four diagnostic levels.

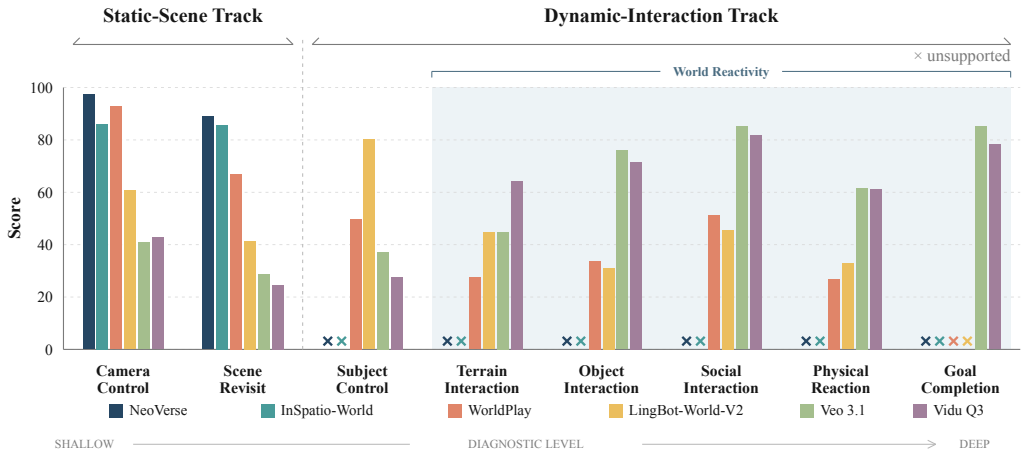


Figure 16: **Task-level profiles.** Two models per interface illustrate strengths in control, scene revisit, and interaction. Crosses denote unsupported tasks.

F.2 Input Translation Multiplier Ablation

Table 7 provides the complete metric breakdown for the NeoVerse ablation in Section 5.4. The input translation multiplier varies from $0.10\times$ to $2.00\times$ with rotations unchanged. Figure 17 shows the corresponding metric trends. As the multiplier increases, Camera Control decreases from 98.25 to 95.32 and Scene Revisit from 90.60 to 88.06. Revisit Success remains 1.000 throughout, while PSNR decreases from 22.68 to 21.32, LPIPS increases from 0.128 to 0.151, and SSIM decreases from 0.704 to 0.641. The Scene Revisit decline therefore reflects reduced appearance fidelity at successfully recovered viewpoints, rather than an increased frequency of failed returns.

All five general metrics also decline, with Photometric Consistency falling from 80.17 to 62.45 and the General average from 80.42 to 75.05. These changes show that reducing the requested displacement can improve both task and general scores without changing the model. The ablation therefore motivates aligning visible displacement before comparing camera-driven models, so that their scores are evaluated under comparable motion demands (Section 4.1).

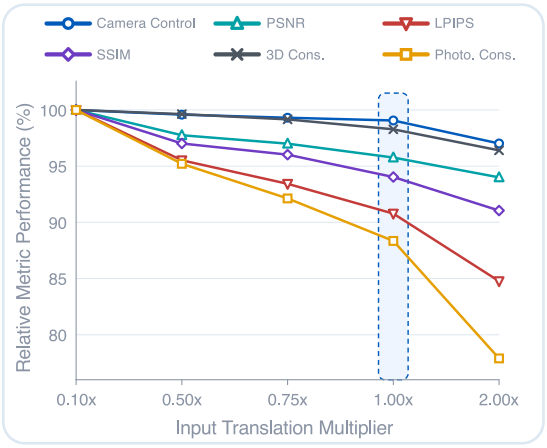


Figure 17: **Input translation ablation.** NeoVerse metrics across five translation multipliers.

Table 7: **Complete translation-multiplier ablation for NeoVerse.** Only input translations vary; rotations are unchanged. Scores are averaged per case.

Multiplier	Camera Control			Scene Revisit					3D Cons.	Photo. Cons.	Temp. Flick.	Aesth. Quality	Imag. Quality	Average		
	$e_t \downarrow$	$e_r \downarrow$	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	90.60	99.91	80.17	95.41	54.06	72.53	94.43	80.42	87.43
0.50×	0.003	0.51	97.84	1.000	22.17	0.134	0.683	89.81	99.54	76.33	94.21	53.39	72.47	93.83	79.19	86.51
0.75×	0.004	0.56	97.57	1.000	22.00	0.137	0.676	89.50	99.07	73.86	93.81	53.03	72.30	93.54	78.41	85.98
1.00×	0.005	0.60	97.33	1.000	21.72	0.141	0.662	88.97	98.19	70.83	93.53	52.72	72.19	93.15	77.49	85.32
2.00×	0.015	0.91	95.32	1.000	21.32	0.151	0.641	88.06	96.32	62.45	92.93	52.00	71.54	91.69	75.05	83.37

F.3 Sensitivity to the Reconstruction Backend

We replace VGGT-Ω with Depth Anything 3 (DA3) (Lin et al., 2025) while keeping the benchmark inputs and generated videos fixed. Tables 8 and 9 report the results for all 23 static-scene models and all 12 dynamic-interaction models, respectively.

Table 8: **Static-scene track evaluation with DA3.** Task and General average the two task scores and five general metrics; Overall averages both. Higher is better except e_t, e_r, and LPIPS. Bold/underlined values mark the best/second-best within each interface group.

Model	Camera Control			Scene Revisit					3D Cons.	Photo. Cons.	Temp. Flick.	Aesth. Quality	Imag. Quality	Average		
	$e_t \downarrow$	$e_r \downarrow$	Score↑	Success↑	PSNR↑	LPIPS↓	SSIM↑	Score↑						Task	General	Overall
Camera-driven																
TrajectoryCrafter (Yu et al., 2025)	0.13	1.10	81.97	**1.000**	19.06	0.240	0.544	82.81	95.63	62.83	92.56	52.18	67.03	82.39	74.05	78.22
ReCamMaster (Bai et al., 2025)	0.47	4.01	38.19	0.815	16.03	0.352	0.415	62.02	**99.18**	**81.97**	95.51	54.58	73.45	50.11	**80.94**	65.53
Voyager (Huang et al., 2025a)	0.30	4.04	60.03	0.995	16.89	0.386	0.467	75.80	89.81	27.44	93.04	53.59	63.69	67.92	65.51	66.72
FantasyWorld (Dai et al., 2026)	1.05	7.76	18.24	0.555	15.39	0.323	0.373	41.88	98.79	75.60	**95.90**	**57.11**	**73.98**	30.06	80.28	55.17
NeoVerse (Yang et al., 2026)	**0.01**	**0.72**	**96.93**	**1.000**	**21.70**	**0.141**	**0.661**	**88.92**	97.53	70.83	93.53	52.72	72.19	**92.93**	77.36	**85.15**
InSpatio-World (1.3B) (Team et al., 2026a)	0.08	1.21	86.91	**1.000**	20.77	0.225	0.605	85.58	97.79	64.88	93.68	53.77	73.60	86.25	76.74	81.50
Action-driven																
Hunyuan-GameCraft (Li et al., 2025)	0.13	11.92	39.17	0.375	13.37	0.542	0.369	28.36	94.00	53.93	93.91	54.98	71.70	33.77	73.70	53.74
Astra (Zhu et al., 2026b)	0.71	7.97	15.97	0.360	12.18	0.600	0.309	23.38	97.09	78.42	96.37	51.71	71.78	19.68	79.07	49.38
WorldPlay (Sun et al., 2025b)	**0.08**	**1.06**	**88.31**	0.780	**18.69**	**0.269**	**0.544**	66.66	98.55	81.06	95.90	53.92	73.31	**77.49**	80.55	**79.02**
Yume 1.5 (Mao et al., 2026)	0.32	3.14	59.69	0.270	12.37	0.614	0.359	17.52	97.74	67.53	95.18	53.42	75.27	38.61	77.83	58.22
LingBot-World (Team et al., 2026b)	0.19	6.28	54.43	0.620	15.63	0.340	0.409	47.20	99.42	81.52	**96.53**	**59.87**	75.01	50.82	82.47	66.65
Infinite-World (Wu et al., 2026c)	0.25	2.38	64.49	0.655	14.30	0.389	0.345	47.10	**99.87**	**92.09**	95.99	56.27	**76.40**	55.80	**84.12**	69.96
Matrix-Game 3.0 (Wang et al., 2026b)	0.17	4.27	68.93	0.860	13.83	0.486	0.372	60.26	95.61	30.14	93.39	48.80	73.39	64.60	68.27	66.44
SANA-WM (Zhu et al., 2026a)	0.09	1.94	83.60	**0.910**	16.37	0.365	0.449	**68.91**	99.04	75.50	95.36	54.47	73.43	76.26	79.56	77.91
LingBot-World-V2 (Gao et al., 2026)	0.23	7.65	54.48	0.585	12.77	0.514	0.314	39.68	92.36	46.73	92.50	55.12	71.55	47.08	71.65	59.37
ABot-World (Jiang et al., 2026)	0.37	5.43	49.31	0.860	11.93	0.664	0.371	54.25	83.32	10.30	95.38	47.31	65.64	51.78	60.39	56.08
Language-driven																
Kling 2.5 (Kuaishou Technology, 2025)	0.31	6.93	43.37	0.315	13.50	0.535	0.380	22.81	**99.78**	**87.87**	**97.57**	56.28	75.70	33.09	**83.44**	58.27
Veo 3.1 (Google DeepMind, 2025)	0.89	8.20	21.27	0.447	12.30	0.595	0.345	28.77	98.73	73.61	95.42	55.49	76.57	25.02	79.96	52.49
Hailuo 2.3 (MiniMax, 2024)	**0.28**	**5.18**	**54.61**	**0.510**	**13.56**	**0.518**	**0.388**	**36.87**	98.94	79.88	95.32	**56.30**	75.52	**45.74**	81.19	**63.47**
Wan 2.6 I2V (Wan et al., 2025)	0.38	6.12	46.50	0.485	13.29	0.527	0.355	34.65	99.27	69.43	94.13	53.93	**77.53**	40.58	78.86	59.72
Seedance 1.5 (Seedance et al., 2025)	0.78	7.78	28.47	0.370	12.29	0.601	0.343	23.40	95.96	56.60	94.82	54.45	74.89	25.94	75.34	50.64
Vidu Q3 (Bao et al., 2024)	0.49	8.48	30.69	0.360	12.67	0.582	0.352	24.54	99.13	71.27	95.01	55.38	76.73	27.62	79.50	53.56
HappyHorse 1.0 (Alibaba Group, 2026)	0.40	5.55	45.66	0.415	12.45	0.564	0.351	28.42	99.52	74.17	95.02	55.57	76.99	37.04	80.25	58.65

Table 9: **Dynamic-interaction track evaluation with DA3.** Task and General average supported task scores and four general metrics; Overall averages both. Higher is better. Dashes denote unsupported tasks. Bold/underlined values mark the best/second-best within each interface group.

Model	Subject Control	Terrain Inter.	Object Inter.	Social Inter.	Physical Reaction	Goal Comp.	Subject Cons.	Motion Smooth.	Aesth. Quality	Imag. Quality	Average		
											Task	General	Overall
Action-driven													
WorldPlay (Sun et al., 2025b)	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 (Team et al., 2026b)	53.43	21.63	25.94	**60.37**	33.43	–	**94.98**	**98.89**	60.87	**71.70**	38.96	**81.61**	60.29
SANA-WM (Zhu et al., 2026a)	65.53	28.90	**40.19**	48.71	36.41	–	92.37	98.55	55.79	71.28	43.95	79.50	61.72
LingBot-World-V2 (Gao et al., 2026)	**77.42**	38.29	31.24	45.71	32.83	–	89.27	97.99	**61.39**	71.36	45.10	80.00	62.55
ABot-World (Jiang et al., 2026)	66.44	**50.19**	36.88	48.08	**39.17**	–	87.59	98.26	57.54	67.72	**48.15**	77.78	**62.96**
Language-driven													
Kling 2.5 (Kuaishou Technology, 2025)	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 (Google DeepMind, 2025)	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 (MiniMax, 2024)	**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 (Wan et al., 2025)	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 (Seedance et al., 2025)	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 (Bao et al., 2024)	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 (Alibaba Group, 2026)	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