WHAT-IF WORLD: A CONTRASTIVE REASONING BENCHMARK FOR VIDEO WORLD MODELS IN EMBODIED SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation models are increasingly used as world simulators for driving and robotic manipulation, where their value lies in answering what-if questions: if the car brakes harder, or the object is heavier, what happens next? Answering such questions requires a model to change its output correctly when its input changes. Many existing benchmarks score each video in isolation rather than test this contrast: a model can generate two videos that each look plausible yet are wrong relative to each other. We introduce What-If World, a contrastive reasoning benchmark of 319 prompt pairs grounded in real frames from nuScenes and DROID. The two prompts in each pair describe the same scene with a single physical variable altered, drawn from a taxonomy of six variables shared across driving and manipulation. The prompts differ only slightly in wording, but the correct videos should diverge substantially. We score each pair with APEO, a rubric with four checks: whether each video follows its prompt (Adherence), whether its motion is physically consistent (Physics), whether the shared scene is preserved (Environment), and whether the two videos differ in the expected outcome (Outcome). Across nine state-of-the-art models, contrastive reasoning remains challenging: none exceeds 53% on the paired score, and the four open-source models average 28.6%. This failure is easy to miss when each video looks plausible on its own. Every model scores 28 to 49 points lower on paired physics than on single-video physics, a gap we call the contrastive bottleneck: a common failure is to produce similar trajectories under both prompts. Beyond being hard to see, the failure falls on the what-if questions that matter for embodied agents; for example, models are often far better at making things move than at making them stop. Realism therefore does not imply contrastive reasoning, which What-If World measures one intervention at a time, a useful test before a video generation model serves as a world simulator.

1 INTRODUCTION

Video generation models are increasingly framed as general-purpose world simulators for embodied applications such as autonomous driving and robotic manipulation (Google DeepMind, 2025; OpenAI, 2024). In these settings, visual realism is typically not enough, and neither is isolated physical plausibility (Motamed et al., 2026; Kang et al., 2025). A model can render a car braking hard with convincing kinematics yet produce nearly identical footage when the prompt is changed to a gentler brake. It reproduces the visual template of "a car braking" but remains insensitive to the input that should shape the outcome. For action-conditioned simulators, model-based planners, and policy training pipelines, this is a critical failure, since rollouts that fail to track the controlled variable are unreliable for comparing alternative actions or conditions.

This failure is comparative: a single video does not reveal how the model responds to a different input. Yet video-quality benchmarks such as VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024b), together with physics-focused extensions like VideoPhy (Bansal et al., 2024) and PhyGenBench (Meng et al., 2025), score each generation against a single prompt. Favorable individual scores for the two braking videos do not establish whether their contrast is correct. Reasoning benchmarks for visual understanding, including CLEVRER (Yi et al., 2020) and NExT-QA (Xiao

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.

et al., 2021), instead test whether a model perceives causal structure in a video shown to it—not whether the videos it generates track changes in its input. These evaluations do not directly test whether a controlled change in input yields a correspondingly controlled change in output. We call this capability contrastive reasoning.

Probing this capability in video generation models requires answering three questions: what to test, how to test, and how to evaluate. First, there is no agreed taxonomy of physical intervention variables for embodied video generation, so the test space must be defined before any test can be selected. Second, controlled comparison is fragile in a generative setting: the model is text-driven yet must produce a physical effect, and *ceteris paribus* must be approximated through input design rather than enforced through generation. Third, evaluation must distinguish omitted interventions, insufficient divergence, and divergence caused by scene drift. These judgments require semantic checks beyond generic feature distances, while full human annotation is impractical at this scale.

To fill these gaps, we build **What-If World**, a benchmark for contrastive reasoning in autonomous driving and robotic arm manipulation. To define what to test, two domain experts apply three admission criteria (physically fundamental, embodiment-shared, operationally isolable) to real videos and merge survivors by physical mechanism (e.g., braking and acceleration intensity collapse into a single Force/Degree primitive). This yields six primitives in two groups: environment (surface friction, material/medium, obstacle configuration) and agent action (spatial alignment, force/degree, temporal sequencing). To decide how to test, we build each instance in two stages. (1) We filter clips from nuScenes (Caesar et al., 2020) or DROID (Khazatsky et al., 2024) for the *physical conditions* the target intervention needs, extract the frame just before the action begins, and pair it with two scene-describing prompt components (camera perspective and initial scene state), which together anchor the shared initial state. (2) We author a prompt pair that differs in a single physical variable. To evaluate, we introduce the **APEO** framework, scoring each pair on Adherence (intervention executed), Physics (dynamics valid), Environment (shared scene preserved), and Outcome (result diverged as predicted); the first three are scored in both single-video and paired modes, while Outcome is paired-only. The rubric runs at benchmark scale via a VLM judge validated against human annotation.

Applying this protocol to nine state-of-the-art video generation models (four open-source, five closed-source) with Gemini 3.1 Pro as the VLM judge, we report three findings. (1) **Contrastive reasoning is far from solved.** No model exceeds 53% on our paired metric, and the four open-source models average 28.6%. (2) **Per-video realism hides a contrastive bottleneck.** A common failure is to produce similar trajectories under both prompts. Paired physics falls 28 to 49 points below single-video physics for all nine models, a gap we call the contrastive bottleneck. This failure is not directly measured by single-video physics scores, and paired scoring also reveals differences between models: the three leading models are within 0.9 points of each other on single-video scores but 8.7 points apart on the paired score. (3) **Models often fail on interventions that embodied agents rely on.** For example, models are often far better at making things move than at making them stop: in the same car-following scenes, every model gets acceleration right more often than braking (54.7% against 24.9%; 92.6% against 11.1% for the best model), and on a snowy road a car often brakes just as it would on dry asphalt. Realism therefore does not imply contrastive reasoning, and a video model should be tested intervention by intervention before serving as an embodied simulator.

In summary, our contributions are as follows.

1. **The What-If World benchmark**, a benchmark unifying physical interventions in driving and manipulation through six primitives and 319 prompt pairs anchored on real frames from nuScenes and DROID.
2. **The APEO evaluation framework**, a protocol scoring Adherence, Physics, and Environment in both single-video and paired modes, and Outcome on pairs, with a VLM-based implementation validated against human annotation.
3. **A diagnostic study of nine state-of-the-art video generation models**, which shows that contrastive reasoning is far from solved, identifies the contrastive bottleneck, and finds that models often fail on interventions that embodied agents rely on, such as braking or a slippery road.

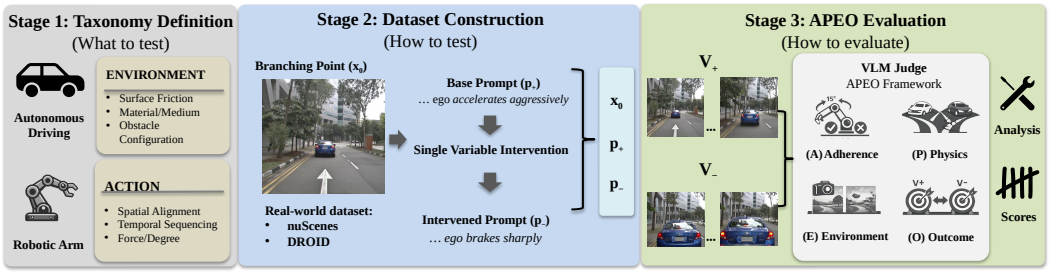


Figure 1: What-If World benchmark overview. Three stages address the challenges of evaluating the contrastive reasoning of video world models: (1) what to test, (2) how to test, and (3) how to evaluate.

2 Background and Related Work

Video generation benchmarks primarily assess individual generations. Existing benchmarks for video generation measure a wide range of properties. One line of work focuses on visual quality and faithfulness (Huang et al., 2024; Sun et al., 2025; Zheng et al., 2025; Huang et al., 2025). Others assess controllability, quality and dynamics (Duan et al., 2025), instruction following and physical plausibility in individual clips (Bansal et al., 2024; Meng et al., 2025; Bansal et al., 2025; Li et al., 2025), or spatial–temporal coherence (Yuan et al., 2024). More recently, a separate line of work evaluates video world models on embodied downstream tasks such as autonomous driving and robotics (Zhou et al., 2026; Qin et al., 2024; Shang et al., 2026). These benchmarks do not directly target the paired comparison studied here: whether a model's output changes, in the direction physics predicts, when one physical variable in the prompt is changed. This gap matters because visual realism is a poor proxy for physical correctness (Motamed et al., 2026; Kang et al., 2025): current generators routinely produce convincing footage that violates the underlying physics, undermining the physical fidelity that embodied applications require (Gupta et al., 2024; Wang et al., 2025a).

Many reasoning benchmarks focus on understanding. VQA benchmarks have been used to evaluate the ability of models to understand causal or physical phenomena (Yi et al., 2020; Xiao et al., 2021; Grunde-McLaughlin et al., 2021). The recent evaluation trend has shifted toward large language models (Wang, 2024; Chi et al., 2024; Maasch et al., 2025) and vision-language models (Wu et al., 2024; Foss et al., 2025; Parmar et al., 2024; Komanduri et al., 2025; Li et al., 2024); these benchmarks pose causal questions, over varied visual inputs in the VLM case, with answer formats including labels and open-ended responses. Moreover, image editing benchmarks (Zhang et al., 2023) take a real reference photograph and measure whether a requested image edit has been applied correctly—no temporal dynamics are required, and the evaluation compares the output to the original. However, with the recent emergence of generative world models (Hafner et al., 2023; Alonso et al., 2024; Wang et al., 2024; Gao et al., 2024b), neither paradigm directly tests whether generated dynamics respond correctly to a physical intervention. Counterfactual dynamics prediction is also studied in CoPhy and Filtered-CoPhy (Baradel et al., 2020; Janny et al., 2022) in synthetic physical systems. What-If World focuses on real embodied scenes: starting from a shared real scene context, the model must generate two videos that differ as the physical intervention predicts.

3 The What-If World Benchmark

What-If World aims to benchmark a specific capability: whether a video generation model can simulate the differential physical consequences of a controlled intervention, rather than merely render visually plausible footage. Formally, a model $\mathcal{M}$ takes an initial frame x_0 and a text prompt p to generate a video $\mathcal{V} = \mathcal{M}(x_0, p)$. We call the capability under test *contrastive reasoning*: given a shared scene context and paired inputs $(x_{0,+}, p_+)$ and $(x_{0,-}, p_-)$ that differ only in the target physical variable v, the model $\mathcal{M}$ should produce two videos $(\mathcal{V}_+, \mathcal{V}_-)$ that diverge in the direction physics predicts under the intervention. We borrow the notion of an intervention from causal inference as a design principle for building such pairs. Building the benchmark answers three questions: what to test (Section 3.1), how to test (Section 3.2) and how to evaluate (Section 3.3), as outlined in Figure 1.

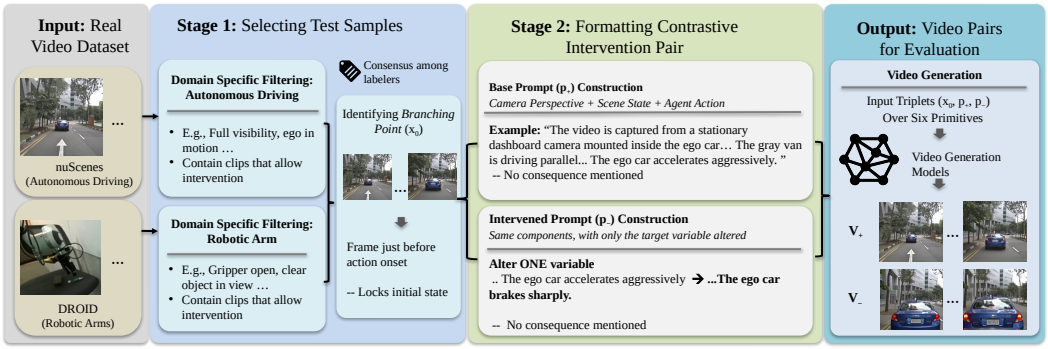


Figure 2: Benchmark construction pipeline. Stage 1 filters clips from nuScenes and DROID and extracts the branching point x_0 at action onset. Stage 2 writes a contrastive prompt pair (p_+, p_-) that shares scene context and differs only in one physical variable, with the outcome withheld. Each paired input is fed to a video model to produce $(\mathcal{V}_+, \mathcal{V}_-)$.

3.1 The What-If World Taxonomy

To address the **What to test** challenge, we need an explicit set of variables that govern physical outcomes. Such a set makes "changing one physical variable" operational, and it lets each failure be traced to a class of physical phenomena rather than to a single opaque "physics" score.

We build this set from real data. Two domain experts (one in driving, one in manipulation) reviewed 1,000 randomly selected videos from nuScenes (Caesar et al., 2020) and DROID (Khazatsky et al., 2024) following a thematic-analysis protocol (Braun & Clarke, 2006) to list candidate variables, and retained one only if it was (i) *physically fundamental*, driven by a physical mechanism rather than a stylistic or perceptual change (Pearl, 2009; Kang et al., 2025); (ii) *embodiment-shared*, applicable to both driving and robotics, to cover related physical effects in both domains (O'Neill et al., 2024; Foss et al., 2025); and (iii) *operationally isolable*, so that it can be changed while everything else is held fixed (Ahmed et al., 2020; Schölkopf et al., 2021). Candidates that act through the same mechanism were merged (Appendix D.2), leaving six primitives in two groups: properties of the environment ($\mathcal{D}_{\text{env}}$) and parameters of the agent's action ($\mathcal{D}_{\text{act}}$).

$\mathcal{D}_{\text{env}}$ — Environment. External conditions that change what a given action produces. (1) *Surface Friction* determines whether contact yields traction or slip: hard braking decelerates cleanly on dry asphalt but skids on ice. (2) *Material/Medium* sets the resistance and deformability of the bodies and the medium they move through, which shapes how force translates into motion, so that the same gripper closure holds a rigid block but crushes a sponge. (3) *Obstacle Configuration* covers whatever in the scene blocks the intended motion, such as a cone in the driving lane or an object too heavy for the arm to push. Target-object mass falls in this category, whereas action intensity falls under Force/Degree (Appendix D.5).

$\mathcal{D}_{\text{act}}$ — Action. Parameters of the action itself, holding the environment fixed. (1) *Spatial Alignment* is where the action is applied: small positional errors separate a clean merge from a sideswipe, or a grasp from a miss. The same primitive covers the direction of longitudinal motion (accelerating to close a gap or braking to open it), whereas Force/Degree covers how hard the same action is performed. (2) *Force/Degree* is how hard or how far (force, speed, or angular extent), and scales the outcome continuously until it crosses a categorical boundary, such as stopping in time versus not, or holding versus dropping. (3) *Temporal Sequencing* is when each sub-action fires relative to the others; closing the gripper before rather than after reaching the target turns the same motion into a grasp or a miss.

Table 14 in Appendix D instantiates each primitive in both driving and manipulation. The two instantiations share the physical variable but not the task, so we also report results for driving and manipulation separately (Appendix G.3).

3.2 BENCHMARK CONSTRUCTION

To address the **How to test** challenge, we design each test instance as a controlled probe of the model's world-simulator capability: from a shared scene context, a change to one physical variable should drive the simulated rollout to diverge as physical law predicts. Two requirements (Pearl, 2009; Schölkopf et al., 2021) make this probe interpretable. The first is state fixation: $\mathcal{V}_+$ and $\mathcal{V}_-$ should originate from the same initial scene apart from v, so that other initial differences do not confound the contrast. The second is single-variable change: only one key variable should differ between the two prompts, so that divergence can be traced to a specific mechanism. We approximate both requirements at the input level with a two-stage construction (Figure 2): (1) selecting a real clip and extracting a frame x_0 that, together with a fixed scene-state description, anchors the initial state, and (2) writing a paired prompt (p_+, p_-) that differs only in the target variable.

Anchoring the state: x_0 at the branching point. We anchor the initial state through a paired specification: a single image frame x_0 together with the *camera perspective* and *initial scene state* components of the prompt. The frame anchors visible geometry, appearance, agent positions, and lighting, while the two prompt components supply information that a single frame cannot encode: how the camera will behave over time (e.g., *"stationary dashboard camera"*) and dynamic context such as the ego vehicle's speed or its motion relative to other agents. This split mirrors how current video generation models are conditioned (image as visual anchor, text as the source of non-visual initial information), and the three together form the shared initial conditioning that we hold fixed across $\mathcal{V}_+$ and $\mathcal{V}_-$, apart from the target variable.[1]

Frames are drawn from nuScenes (Caesar et al., 2020) (autonomous driving) and DROID (Khazatsky et al., 2024) (robotic arm manipulation). Within these datasets, a clip is eligible only if it carries the *physical affordances* that make a target primitive testable, such as the ego car driving side-by-side with another vehicle in an overtaking scenario for *Temporal Sequencing*, or an unoccluded gripper near a target object for manipulation primitives. Domain-specific visibility filters further remove parked ego vehicles, heavy occlusion, and gripper states in or after contact, so that any divergence between $\mathcal{V}_+$ and $\mathcal{V}_-$ remains visible against an uncluttered scene.

From each qualified clip we take the frame immediately preceding action onset as x_0 and call this moment the *branching point*: $\mathcal{V}_+$ and $\mathcal{V}_-$ can start from the same state, and the intervention has not yet taken effect, so it can still change the future. In the overtaking scenario, for instance, x_0 is the moment the ego car and an adjacent vehicle travel side by side at the same speed, where the future trajectory depends on the subsequent action (Appendix D.3).

Isolating the variable: the contrastive prompt pair. Since text is the primary control input that current video generation models accept, we implement the contrastive design in the prompt pair. With the camera perspective and the rest of the scene fixed, the intervention prompt p_- is obtained by editing only the span that describes v: the agent's action for action primitives, and the changed scene property for environment primitives (with an edited frame for 72 pairs). Hence p_+ and p_- differ only in how they describe the target variable. For example, "*[stationary dashboard camera]... a gray van drives parallel to the ego car at the same speed; after a short moment, the ego car accelerates aggressively.*" In p_- the last clause becomes "*the ego car brakes sharply.*" Both prompts withhold the physical consequence. A prompt states how the action is performed, including how strongly (e.g., a gentle crawl versus a maximum-effort launch), but never the effect we evaluate: neither prompt states whether the ego car ends up ahead or behind, how far it travels, or how dark a trace becomes. The model therefore has to infer the outcome from x_0 and v rather than read it from the text, which is the capability the benchmark probes. The full prompt-building process is in Appendix D.4. The benchmark contains 319 pairs in 18 sub-categories, 190 in driving and 129 in manipulation (statistics in Appendix Table 15).

3.3 EVALUATION METRICS

To address the **How to evaluate** challenge, a metric must check two things. Each video must be a valid generation on its own, executing its action, obeying physics and keeping its scene stable. Otherwise, differences between $\mathcal{V}_+$ and $\mathcal{V}_-$ may reflect generation failures rather than the intended

[1] For 72 environment pairs (e.g., snow or an added cone), one prompt is paired with an edited copy of the frame that shows the change, to ensure that each prompt is plausible given its frame (Appendix D.3).

intervention. This is what prior video benchmarks measure. But two individually plausible videos can still ignore v, producing identical or randomly different rollouts, and detecting this requires comparing them jointly, as paired designs do for vision–language models (Thrush et al., 2022; Goyal et al., 2017; Gardner et al., 2020).

Table 1: **The APEO evaluation framework.** Each check is scored 1 (pass) or 0 (fail). A, P and E are scored in both single-video and paired modes; O is paired only.

Dimension	Single-video check (s): pass if	Paired check (p): pass if
A *Adherence*	$\mathcal{V}$ shows the intervention prescribed by p (the action, or the changed condition) on the intended target entity.	The interventions rendered in $\mathcal{V}_+$ and $\mathcal{V}_-$ differ visibly, and the difference follows the direction the prompts specify.
P *Physics*	Motion within $\mathcal{V}$ respects basic physical constraints throughout (no teleportation, morphing, or ghost forces).	$\mathcal{V}_+$ and $\mathcal{V}_-$ follow the same trajectory before the intervention and diverge after it as physics predicts under the prescribed change.
E *Environment*	Background, camera viewpoint, and non-target object permanence remain stable throughout $\mathcal{V}$.	The backgrounds of $\mathcal{V}_+$ and $\mathcal{V}_-$ are near-identical; any visible difference between them comes from the intervention, not from scene artifacts.
O *Outcome*	*Not applicable.* (Outcome is defined only for a pair.)	$\mathcal{V}_+$ and $\mathcal{V}_-$ differ measurably in state at some point (not necessarily the end), in the direction and on the entities that v predicts.

The APEO framework. Given a pair, what should $\mathcal{V}_+$ and $\mathcal{V}_-$ be compared on? We use four dimensions that together cover the conditions a pair must meet to reflect the physical effect of v: Adherence (the intervention was carried out), Physics (the dynamics are valid), Environment (the shared scene is preserved) and Outcome (the result differs in the predicted direction). These checks relate to downstream uses: Adherence to training action-conditioned policies and world models (Kim et al., 2024; Bruce et al., 2024), Physics to inverse and latent dynamics learning (Baker et al., 2022; Hafner et al., 2023), Environment to using generated videos as training data (Hu et al., 2023), and Outcome to planning, which compares futures under different actions (Hafner et al., 2023) (Appendix E.2).

A, P and E are checked both per video and per pair (Table 1). The single-video checks (A_s, P_s, E_s) assess each generation but do not directly score whether the pair diverges correctly; the paired checks (A_p, P_p, E_p, O_p) compare the two videos under *ceteris paribus*. Within paired evaluation, P_p checks how the trajectories diverge over time, whereas O_p checks whether the predicted state difference appears; a correct state difference alone does not ensure a consistent trajectory. A_p, P_p and E_p partly overlap with O_p but separately identify where a pair fails: the outcome can be wrong because the intervention was not rendered (A_p), the dynamics broke (P_p) or the scene drifted (E_p).

The two modes catch different failures. A model can pass the single-video physics and environment checks yet fail P_p and O_p if both videos are individually plausible but follow the same trajectory, that is, if the model ignores v (Section 4.2). The paired checks also guard one another: a pair can pass O_p for the wrong reason, diverging through hallucinated scene artifacts rather than through v, a failure E_p is designed to catch.

VLM-based evaluation with human verification. Manual annotation at this scale is impractical, so we use a VLM judge (Gemini 3.1 Pro), as is now common in video generation benchmarks (Bansal et al., 2024; Li et al., 2025), consistent with the broader LLM-as-a-Judge literature (Zheng et al., 2023; Liu et al., 2023). For each APEO check, the judge receives $(\mathcal{V}_+, \mathcal{V}_-)$ along with (p_+, p_-) and answers a sub-category-specific binary question (e.g., "Does $\mathcal{V}_-$ exhibit a shorter stopping distance than $\mathcal{V}_+$?"). We use binary scoring to make the pass/fail criteria explicit while still aggregating to fine-grained scores through dimension- and instance-level averaging. We validate against 421 human-annotated samples: the judge agrees with human labels on 82.5% of decisions (Cohen's κ = 0.65), comparable to the inter-human agreement of 84.1% (κ = 0.67). Per-check and per-primitive agreement, a second-judge comparison and control experiments are in Appendix E.7.

Table 2: **What-If World leaderboard** (pass rates in %, pooled over three runs). s: single-video; p: paired. Best, second and third results are in red, orange and blue. * closed-source.

Model	A		P		E		O	APEO		Rk.
	s	p	s	p	s	p	p	sAvg	pAvg	
CogVideoX1.5-5B	36.2	19.2	45.4	12.1	27.6	51.6	11.9	36.4	23.7	9
Wan2.2-5B	39.8	22.5	54.6	15.4	48.4	62.1	15.7	47.6	28.9	7
HunyuanVideo1.5-8.3B	41.8	21.0	63.1	13.8	49.9	64.7	11.1	51.6	27.6	8
Cosmos-Predict2-2B	42.4	25.5	57.9	20.2	58.0	75.5	15.5	52.8	34.2	5
Seedance 1.5*	49.7	30.1	56.5	19.0	42.1	46.9	23.9	49.4	30.0	6
Kling 3.0*	56.7	35.6	68.2	26.3	67.4	76.6	22.9	64.1	40.4	4
Seedance 2.0*	62.5	43.1	73.8	31.5	73.9	71.3	31.0	70.1	44.2	3
Veo 3.1*	67.7	50.4	70.2	42.5	69.6	66.8	46.0	69.2	51.4	2
Grok Imagine*	67.1	50.3	71.8	42.3	70.1	74.0	45.1	69.6	52.9	1

4 Experimental Results

We evaluate nine video generation models spanning both open-source and closed-source tiers. The **open-source** models include HunyuanVideo1.5 (Wu et al., 2025), Cosmos-Predict2 (NVIDIA, 2025), CogVideoX1.5 (Yang et al., 2024), and Wan2.2 (Wan Team et al., 2025). The **closed-source** models include Grok Imagine (xAI, 2026), Veo 3.1 (Google, 2025b), Seedance 2.0 (Seedance Team et al., 2026), Seedance 1.5 (Seedance Team et al., 2025), and Kling 3.0 (Kuaishou Technology, 2026). All models receive the same benchmark inputs, including branch-specific frames for the edited environment pairs; no model is fine-tuned or adapted for this benchmark. Each test instance yields one video per prompt ($\mathcal{V}_+$ and $\mathcal{V}_-$), and all scores are pass rates (%). Results aggregate three independent generation runs; run variation and bootstrap intervals are reported in Appendices F.2 and F.3.

4.1 Contrastive Reasoning Is Far from Solved

Table 2 reports the full APEO scorecard. Single-video scores (s) rate $\mathcal{V}_+$ and $\mathcal{V}_-$ separately and average the two; paired scores (p) judge the two videos together under the ceteris-paribus constraint. Outcome (**O**) is scored only on pairs. The last two columns average these checks: s**Avg** = mean(A_s, P_s, E_s) summarizes single-video quality, and p**Avg** = mean(A_p, P_p, E_p, O_p) is our primary metric.

Three findings emerge. (1) **Even the best models fail on most pairs.** The best model, Grok Imagine, reaches only 52.9% pAvg, and Veo 3.1 is effectively tied at 51.4% (Appendix F.2). Even these two pass all four paired checks on fewer than 30% of pairs (28.9% and 26.9%), and the open-source models on only 5 to 11% (Appendix Table 27). (2) **The closed-source tier leads.** Closed-source models average 43.8% pAvg against 28.6% for open-source models (Appendix F.3), and the lead holds on all six primitives, from 7.5 pp on Temporal Sequencing to 22.2 pp on Force/Degree (Appendix C.1). (3) **The closed-source lead is concentrated in intervention-sensitive checks.** The closed-source lead is 17 to 20 pp on the checks that track the intervention (A_p, P_p, O_p) but only 3.6 pp on keeping the rest of the scene fixed (E_p). Across the nine models, pAvg tracks A_p and P_p closely (Spearman 0.97 and 0.98) but E_p only loosely (0.62). Stronger models are better mainly at showing the intervention and its consequence in the pair. Yet this ability still lags far behind realism (Section 4.2) and is missing for many interventions (Section 4.3).

4.2 The Contrastive Bottleneck

Single-video checks assess each generation; paired checks assess contrastive reasoning, including whether the output changes correctly when the input does. The second property cannot be read off how realistic a single video looks, and every model does far worse on it. (1) **Every model scores much lower on paired physics than on single-video physics.** The gap $P_s - P_p$ ranges from 27.7 pp (Veo 3.1) to 49.3 pp (HunyuanVideo1.5), with every 95% interval far from zero (Appendix Table 27). The gap also persists when single-video physics requires both videos in a pair to pass:

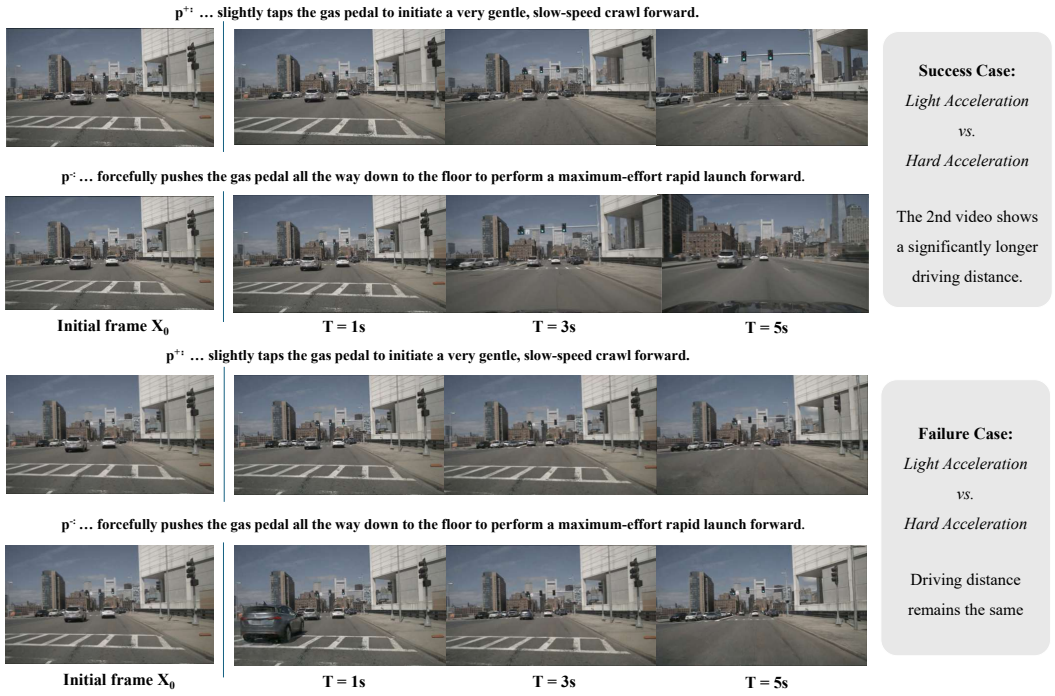


Figure 3: **Success and failure examples (AD).** Both panels start from the same initial frame x_0 and receive a Force/Degree intervention pair that differs only in the acceleration—$\mathcal{V}_+$: "a gentle, slow-speed crawl forward"; $\mathcal{V}_-$: "a maximum-effort rapid launch forward." Frames are sampled at $T = 1, 3, 5$ s. The difference in driving distance is what contrastive evaluation measures.

paired-physics failure rates remain 9.6–35.6 pp higher across the nine models (Appendix Table 5). Thus, the difference is not explained solely by averaging the two single-video scores. We call this gap the *contrastive bottleneck*. At its most extreme, HunyuanVideo1.5 scores 93.9% on single-video physics but 1.1% on paired physics on Overtake Hard (Appendix G.5). (2) **In most paired-physics failures, the model ignores the intervention.** The generated videos often exhibit similar trajectories across the two conditions: in 68.7% of paired-physics failures, the judge's explanations indicate that the two videos follow the same trajectory (Appendix C.2). Figures 3 and 4 show examples. An outcome-hint ablation provides a further check: on the same 269 pairs, explicitly stating the expected outcome raises Grok's O_p from 46.1% to 52.0% and P_p from 45.0% to 47.6% (Appendix C.4). Even with the hint, roughly half of these pairs fail the outcome check. (3) **Single-video scores hide differences between leading models.** Seedance 2.0, Grok Imagine and Veo 3.1 are within 0.9 pp of each other on sAvg, yet span 8.7 pp on pAvg: Seedance 2.0 trails at 44.2%, against 52.9% and 51.4% (Appendix F.2). Overall, the two scores rank the nine models similarly (Spearman 0.90), so single-video scores can separate the tiers, while the paired score further separates these models.

4.3 Where Contrastive Reasoning Is Missing

Averages still hide which interventions a model misses. Broken down by intervention, the paired scores show that models often fail on interventions an embodied agent depends on, such as braking, a slippery road or a heavier object (Appendix C.6). (1) **A model can follow one intervention and miss a closely related one.** On the same 18 car-following scenes (the longitudinal pairs of Spatial Alignment), every model produces the expected outcome more often when told to accelerate than when told to brake: 54.7% against 24.9% on average (95% CI of the gap [23.5, 35.4]), and the best model, Grok Imagine, shows the largest contrast (92.6% against 11.1%). Single-video physics, however, gives the two kinds of video similar scores. (2) **Changes to the world are shown but often have no effect.** Among pairs in which both videos follow their own prompts, the outcome is correct in 80.9% of action pairs but only 45.8% of environment pairs (Appendix C.6). Requiring both videos to pass single-video physics as well preserves the ordering (95.0% against 85.9%; Appendix Table 9).

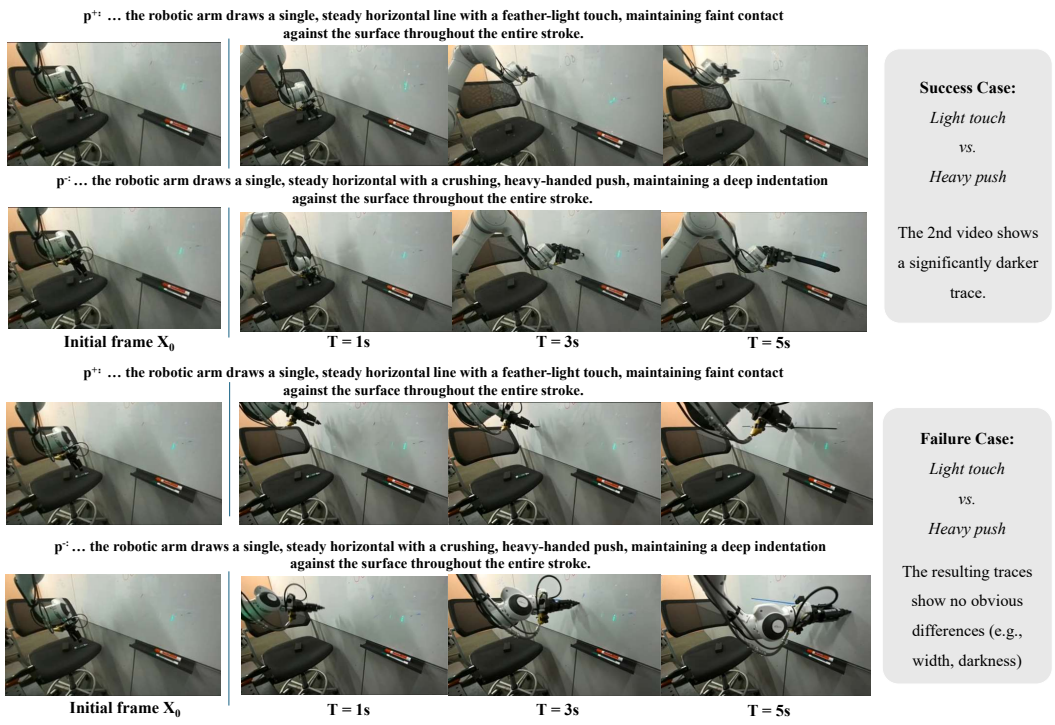


Figure 4: **Success and failure examples (Robotic Arms).** Both panels start from the same initial frame x_0 and receive a Force/Degree intervention pair that differs only in contact intensity—$\mathcal{V}_+$: "with a feather-light touch, maintaining faint contact"; $\mathcal{V}_-$: "with a crushing, heavy-handed push, maintaining a deep indentation." Frames are sampled at $T = 1, 3, 5$ s. The difference in the contact trace (e.g., its darkness and width) is what contrastive evaluation measures.

This compares outcomes among pairs that meet both the adherence and within-video physics criteria. For example, a video may show snow on the road while the car brakes just as it would on dry asphalt (Appendix Figure 6). These results suggest that models treat the environment as appearance rather than as a cause of motion. Braking, friction and mass can each constrain the response to an action, although they act through different physical mechanisms. Such interventions score lower on average than those that add force, and models are often far better at making things move than at making them stop (Appendix C.6). Neither a model's realism nor its success on a related question shows whether it can be trusted on these, so What-If World scores each kind of intervention on its own pairs and identifies missed interventions.

5 Conclusion

We introduced What-If World, a contrastive reasoning benchmark for video world models in embodied settings. Across nine state-of-the-art video generation models, contrastive reasoning remains far from solved: models can produce similar trajectories under both inputs, and they are often far better at making things move than at making them stop. Realism and contrastive reasoning are therefore distinct evaluation targets, and progress on the first does not by itself deliver the second. As video models are increasingly asked what-if questions, they should be judged not only by how plausible their futures look, but by whether those futures change as the world would.

AI USE STATEMENT

We used large language models for editorial assistance, including grammar, wording, consistency and reference checks. All research ideas, experiments, analyses and conclusions are the authors' own, and the authors take full responsibility for the content of this paper. Separately, AI models are part of the benchmark's method and are described where they are used: Gemini 3.1 Pro screens candidate clips, checks prompt pairs, labels failure modes and serves as the VLM judge; Gemini 3 Pro Image edits the initial frames of some environment pairs; and GPT-5 serves as a second judge (Section 3.3 and Appendices D and E).

REFERENCES

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Tayfun Ates, M Ateşoğlu, Çağatay Yiğit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. Craft: A benchmark for causal reasoning about forces and interactions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2602–2627, 2022.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. *Advances in Neural Information Processing Systems*, 32, 2019.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.12000.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

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.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37:96640–96670, 2024.

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.

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

Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12462–12469. IEEE, 2024a.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596, 2024b.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019.

Google. Introducing Nano Banana Pro. `https://blog.google/innovation-and-ai/products/nano-banana-pro/`, 2025a.

Google. Introducing Veo 3.1 and advanced capabilities in Flow. `https://blog.google/innovation-and-ai/products/veo-updates-flow/`, 2025b.

Google DeepMind. Genie 3: A new frontier for world models. `https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/`, 2025. Accessed: 2026-05-06.

Google DeepMind. Gemini 3.1 Pro: Model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11287–11297, 2021.

Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Marc Rigter, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, et al. The essential role of causality in foundation world models for embodied ai. *arXiv preprint arXiv:2402.06665*, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *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*, 2025.

Steeven Janny, Fabien Baradel, Natalia Neverova, Madiha Nadri, Greg Mori, and Christian Wolf. Filtered-CoPhy: Unsupervised learning of counterfactual physics in pixel space. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00368.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *ICML*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Kie.ai. Kie.ai ai model api platform, 2026. URL https://kie.ai/. Unified API for video, image, music and LLM models.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30648–30668, 2025.

Kuaishou Technology. Kling AI launches 3.0 model, ushering in an era where everyone can be a director. https://ir.kuaishou.com/node/11216/pdf, 2026.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. Multimodal causal reasoning benchmark: Challenging vision large language models to discern causal links across modalities. *arXiv preprint arXiv:2408.08105*, 2024.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024a.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

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

Jacqueline RMA Maasch, Alihan Hüyük, Xinnuo Xu, Aditya V Nori, and Javier Gonzalez. Compositional causal reasoning evaluation in language models. *arXiv preprint arXiv:2503.04556*, 2025.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *ICML*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

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

NVIDIA. Cosmos-Predict2-2B-Video2World: Model card. `https://huggingface.co/nvidia/Cosmos-Predict2-2B-Video2World`, 2025.

OpenAI. Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators/`, 2024. Accessed: 2026-05-06.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Paritosh Parmar, Eric Peh, Ruirui Chen, Ting En Lam, Yuhan Chen, Elston Tan, and Basura Fernando. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. *Advances in Neural Information Processing Systems*, 37:92769–92802, 2024.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9856–9870, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Seedance Team, 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.

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

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 *CVPR*, pp. 8406–8416. Computer Vision Foundation / IEEE, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.

Wan Team, 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.

Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. Gensim: Generating robotic simulation tasks via large language models. *arXiv preprint arXiv:2310.01361*, 2023.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the computer vision and pattern recognition conference*, pp. 22442–22452, 2025a.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024.

Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025b.

Zeyu Wang. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 143–151, 2024.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

xAI. Grok Imagine API. `https://x.ai/news/grok-imagine-api`, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pp. 9777–9786. Computer Vision Foundation / IEEE, 2021.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*. OpenReview.net, 2020.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270, 2024.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

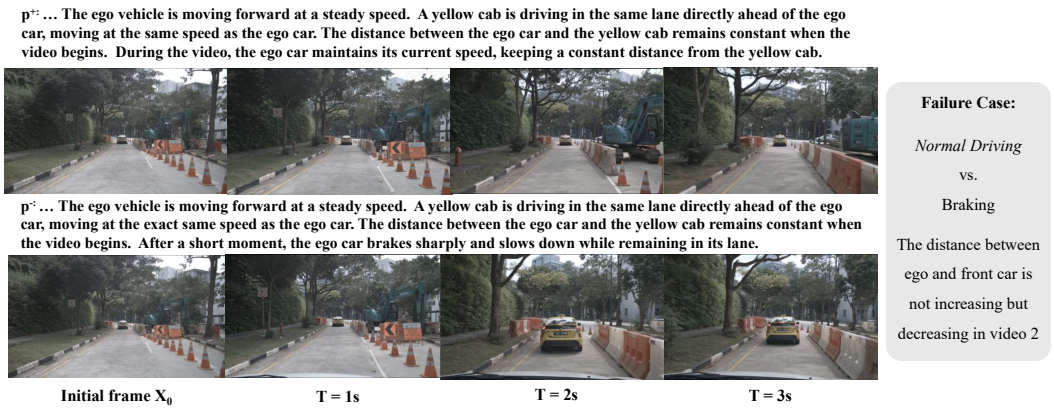


Figure 5: **Reversed outcome (AD, Spatial Alignment).** The ego car is supposed to brake in $\mathcal{V}_-$, which should make the gap to the front car *increase*. Instead, the gap *decreases*—the outcome is reversed. $\mathcal{V}_+$ (constant speed) looks correct on its own.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint arXiv:2402.06119*, 2024.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L Waslander. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. *arXiv preprint arXiv:2601.01528*, 2026.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A Limitations and Future Work

What-If World measures contrastive reasoning rather than downstream utility. Each APEO dimension is anchored to a concrete downstream use (Section 3.3), but empirically validating how these scores translate into deployed performance is left to future work. We also characterize failure modes phenomenologically without attributing them to specific architectural or training-data causes; doing so requires training-side controlled experiments that fall outside the scope of an evaluation benchmark. The benchmark and evaluation code[2] also apply to new models. The benchmark's 319 prompt pairs are comparable in scale to physics-focused video generation benchmarks such as PhyGenBench (160 prompts; Meng et al., 2025) and WorldModelBench (350 image-and-text conditions; Li et al., 2025). With two videos per pair for each model over three generation runs, the closed-source/open-source tier gap and each model's single-to-paired physics gap remain statistically significant under a scene-cluster bootstrap (Appendix F.3).

B Representative Failure Cases – Qualitative Results

Section 4 and Appendix C show *how often* models fail; this section shows *what the failures look like*. For each of four examples, we show the initial frame x_0, the two prompts (p_+ and p_-), and sampled frames from the generated videos. All four cases are from run 1. Each shows one failure mode clearly; Table 4 gives how often each mode occurs.

Reversed outcome in driving (Figure 5). This is a Spatial Alignment test from Autonomous Driving. The two prompts share the same setup (i.e., an ego car following a yellow cab at the same

[2]Anonymized code: `https://anonymous.4open.science/r/What-if-World-388B`. Example data: `https://dataverse.harvard.edu/previewurl.xhtml?token=72495e79-e344-453b-92e9-63d508634b88`.

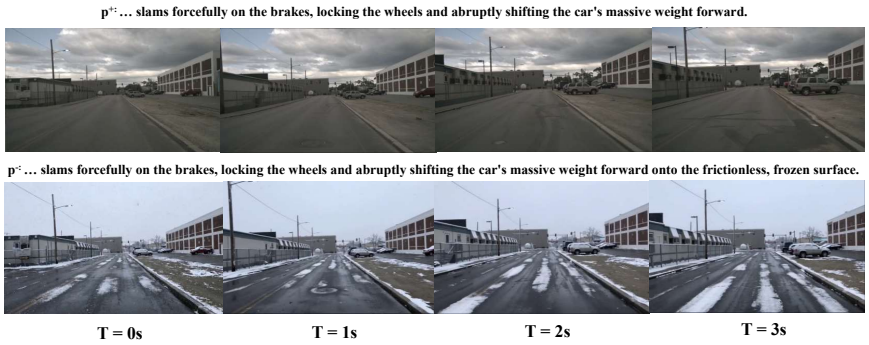


Figure 6: **Mode collapse under friction change (AD, Surface Friction).** Both videos show hard braking, but p_- specifies a frozen, low-friction surface. The snowy road is shown (it comes from the edited initial frame), but the braking distance is nearly the same in both conditions.

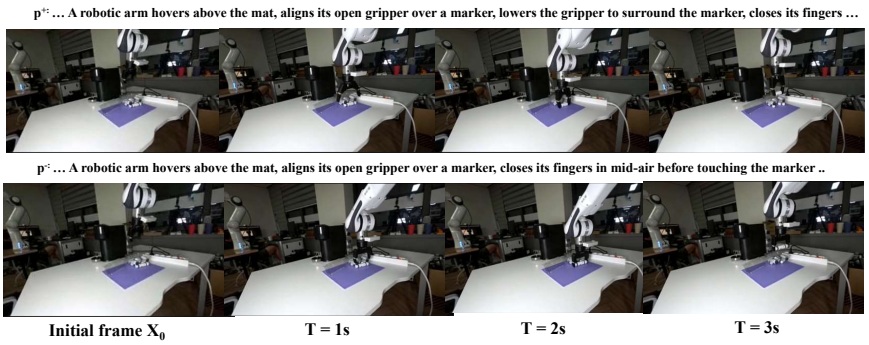


Figure 7: **Mode collapse under timing change (Robotic, Temporal Sequencing).** p_+ closes the gripper after lowering it onto the marker; p_- closes the gripper in mid-air before touching the marker. The premature closure in p_- should cause a miss, but the model picks up the object in both videos—it does not distinguish the two action sequences.

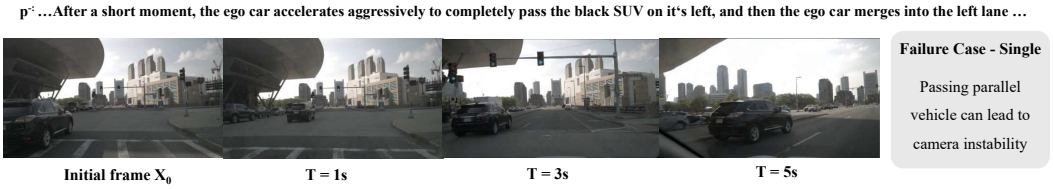


Figure 8: **Single-video camera hallucination (AD, E_s failure).** The ego car accelerates to pass a black SUV and merge left. The model executes the action, but the camera viewpoint drifts severely over time, breaking the stationary dashboard-camera assumption. By $T = 5\,\text{s}$ the perspective no longer matches the original viewpoint.

speed) and differ only in whether the ego car continues at constant speed (p_+) or brakes sharply (p_-). In a correct generation, the gap between the two cars should *increase* in $\mathcal{V}_-$ (because the ego car slows down while the cab keeps going). Instead, the model produces a $\mathcal{V}_-$ where the gap *decreases*—the ego car appears to get closer to the cab despite supposedly braking. This is a textbook reversed outcome (the judge's rubric calls it "causal inversion"): the outcome moves in the opposite direction from what the physics predicts. Meanwhile, $\mathcal{V}_+$ looks reasonable on its own.

Mode collapse under surface friction change (Figure 6). This is a Surface Friction test. Both prompts describe the same hard-braking action, but p_+ takes place on a normal dry road while p_- takes place on a frozen, low-friction surface. On ice, the car should slide much farther before stopping. However, the model produces nearly identical braking behavior in both conditions—the car decelerates at roughly the same rate and covers roughly the same distance. Both videos show their environment (dry road vs. snowy road; the snow comes from the edited initial frame), but the changed surface has almost no effect on the motion. This is mode collapse at the outcome level: the model behaves as if it treated friction as a visual attribute (snow on the ground) rather than a physical property that changes how braking works.

Mode collapse under timing change in manipulation (Figure 7). This is a Temporal Sequencing test from Robotic Manipulation. The two prompts describe the same gripper-over-marker setup, but p_+ instructs the arm to lower the gripper onto the marker *before* closing, while p_- instructs it to close its fingers in mid-air *before* touching the marker. In reality, closing the gripper too early means it cannot grasp anything—it should miss. But the model generates a successful grasp in both videos: the marker gets picked up regardless of timing. The output is consistent with a model that has learned the visual pattern "gripper near object $\rightarrow$ object gets picked up" but does not track that the *sequence* of close-then-lower vs. lower-then-close determines whether the grasp succeeds. Temporal Sequencing is hard even for the best models (Appendix G.1).

Single-video camera hallucination (Figure 8). Unlike the previous three examples, this is a single-video failure (E_s) rather than a paired one. The prompt asks the ego car to accelerate aggressively, pass a black SUV, and merge left. The model executes the overtaking action, but the camera viewpoint drifts dramatically over the course of the video—by $T = 5$ s, the perspective has shifted so far that the scene looks like it was filmed from a different vehicle entirely. This kind of camera instability is a common E_s failure in driving scenarios that involve lateral motion (lane changes, merges), where the model apparently struggles to keep the dashboard-camera viewpoint fixed while generating large lateral displacements.

C Additional Results and Analysis

C.1 Closed-Source vs. Open-Source Gap by Primitive

Table 3 breaks down the closed-source (CS) vs. open-source (OS) performance gap by physical primitive. The overall gap is 15.2 pp, but this number hides substantial variation across primitives.

Force/Degree shows the largest gap (+22.2 pp). This primitive tests whether models can produce different magnitudes of the same action (e.g., gentle vs. hard braking). Closed-source models may benefit from larger capacity and more diverse training data.

Temporal Sequencing shows the smallest gap (+7.5 pp).

Temporal Sequencing remains difficult for both model tiers.

The remaining four primitives fall in a fairly narrow band (about 15 to 21 pp).

C.2 Per-Dimension Failure Analysis

We analyze the VLM judge's reasoning across all evaluated instances of the three runs (about 8,600 in total) to characterize how each APEO dimension fails under single-video and paired evaluation. Table 4 summarizes the dominant failure modes and their prevalence. This table is for analysis only and enters no score or ranking; Appendix C.3 describes how the judge's explanations were labeled and how reliable the labels are.

Table 3: p**Avg (%) by model tier and primitive** (pooled over three generation runs). CS = closed-source (5 models); OS = open-source (4 models). The gap is largest for Force/Degree and smallest for Temporal Sequencing.

Primitive	**CS**	**OS**	**Gap**
Spatial Alignment	49.5	34.6	+14.8
Force/Degree	54.6	32.4	+22.2
Temporal Sequencing	34.5	27.0	+7.5
Surface Friction	34.5	18.3	+16.2
Material/Medium	50.1	29.6	+20.5
Obstacle Configuration	34.8	16.1	+18.7
Overall	**43.8**	**28.6**	**+15.2**

Table 4: **Failure mode taxonomy.** For each APEO dimension, we classify failures into modes based on VLM judge reasoning. Percentages indicate the fraction of failures in that dimension exhibiting each mode (categories can overlap), as the mean of the three generation runs. N = number of failing instances (all three runs together).

Dim.	**Eval**	**Failure Mode**	**%**
A	Single (N=6003)	Action omission (ignores instruction)	50.6
		Action substitution (wrong action)	41.7
	Paired (N=5758)	No discernible inter-video difference	53.5
		Baseline inconsistency (state diverges)	23.8
P	Single (N=4619)	Kinematic discontinuity / teleportation	30.8
		Object / arm morphing	38.4
		Ghost force / levitation	9.9
	Paired (N=6471)	Mode collapse (identical trajectories)	68.7
		No interaction occurs	8.2
E	Single (N=4985)	Object permanence violation	76.3
		Background collapse / distortion	25.2
		Camera hallucination	8.2
	Paired (N=2969)	Cross-video environment divergence	80.1
		Divergent hallucinations	47.0
O	Paired (N=6470)	Mode collapse (same final state)	56.6
		Upstream propagation (action never happened)	37.1
		Reversed outcome (wrong direction)	13.1

Adherence (A). A_s fails when a video does not execute the prompted action. The dominant mode is *action omission* (50.6% of A_s failures): the model generates a plausible driving or manipulation clip while ignoring the instruction—the judge observes *"the ego car maintains constant speed"* when acceleration was prompted, or *"the gripper lowers directly over the target"* when a spatial offset was specified. A secondary mode is *action substitution* (41.7%): the model executes a different action (e.g., the front car accelerates away instead of the ego car accelerating forward). A_p fails differently—it requires that the *difference* between the two videos' actions be visually unambiguous. The most common A_p failure is simply that the two actions look the same (53.5%). The second most common is *baseline inconsistency* (23.8%): both videos individually show plausible actions, but their arm kinematics or vehicle trajectories differ from the very first frame, making it impossible to attribute any divergence to the physical variable rather than to uncontrolled generation variance.

The judge notes: *"the arm approach trajectories are completely different between the two videos—a viewer could not identify which spatial difference is due to the prompt."* (Quoted judge excerpts in this section come from run 1.)

Physics (P). P_s fails when a single video contains internal physical violations—object morphing, teleportation, or impossible kinematics. Typical observations include: *"the robotic arm exhibits severe morphing and loses structural integrity," "the lid distorts into an unrecognizable blob,"* and *"a duplicate object spontaneously spawns mid-video."* Morphing (38.4%) and kinematic discontinuities (30.8%) are the two main modes; ghost forces are rare (9.9%). These are generation-quality failures that have nothing to do with the intervention. P_p failures are fundamentally different: each video may look physically plausible on its own, but their trajectories do not bifurcate at the expected intervention point. The judge mostly reports *mode collapse*—*"the gap closes in both videos at the same rate; there is no trajectory divergence"*—and much less often *no interaction*, where the contact or push that should create the difference never happens in one of the videos (Table 4). Wrong-direction outcomes show up under O_p below, not here. Mode collapse is the most frequent label among paired-physics failures (68.7%): the explanations describe similar trajectories across the two conditions.

Environment (E). E_s fails when the background or non-target objects within a single video are unstable. The judge reports *background collapse* (*"the building on the right liquefies and distorts"*), *object permanence violations* (*"the parked excavator morphs into a moving truck"*), and *unscripted object spawns* (*"a duplicate donut appears on the table while the arm holds another"*). Permanence violations, including spawned and vanishing objects, are by far the most common (76.3%). E_p requires that the two videos share a near-identical "canvas"—same background, same non-target objects, same camera. It fails when generation variance introduces *divergent hallucinations*: *"Video B introduces a green barrier and changes the curb style in areas that should be identical to Video A,"* or *"the workspace background and camera angle are completely inconsistent between the two videos."* Most E_p failures are plain cross-video differences in the scene (80.1%); 47.0% involve one video hallucinating new content.

Outcome (O). O_p measures whether the expected physical state divergence is observable between the pair. Failures fall into three categories. *Mode collapse* (most common, 56.6%): both videos reach the same final state—*"both videos show the cap remaining on the bed; neither grasps it"* or *"the front car appears the same size at the end of both videos; no gap difference emerges." Upstream propagation (37.1%)*: an earlier failure (action omission, object morphing) prevents any meaningful outcome—*"because the arm fails to make contact in either video, the plate does not move in either case." Reversed outcome (13.1%; the judge's rubric calls this "causal inversion")*: the outcome diverges in the wrong direction—*"the acceleration instruction resulted in a widening gap"* or *"the stopping distances are reversed between conditions."* For the action primitives, 91.9% of the O_p failures also fail A_p (the action was not rendered as prompted), against 70.2% for the environment primitives, where more failures render the change but show no effect of it on the motion.

Table 5 reveals an asymmetry across dimensions. **Physics** is the only dimension where paired evaluation is substantially harder than single (+21.5 pp mean Δ): a video that looks physically plausible on its own often does not show the correct *differential* physics when compared with the other video in its pair. **Environment** shows the opposite pattern (−23.4 pp): paired consistency is *easier* than single-video stability, likely because the two videos usually start from the same frame and drift in similar ways, so they stay consistent with each other even as each degrades. **Adherence** shows near-zero Δ (−2.8 pp): whether a model executes the prompted action is equally detectable in single and paired modes.

Model rankings shift across primitives, with no single model dominating. Grok Imagine leads on Force/Degree (68.1% pAvg), Surface Friction (45.4%), and Obstacle Configuration (54.6%); Veo 3.1 leads on Spatial Alignment (61.5%), Temporal Sequencing (42.9%), and Material/Medium (65.9%). These leaders are the same in all three runs, except that Seedance 2.0 leads Surface Friction in run 2. Some specializations are extreme: Grok scores 54.6% on Obstacle Configuration, where the next-best model scores 33.9% (a 21 pp lead). Task-level variance within one primitive is large: Grok scores 92.6% O_p on Longitudinal Acceleration but 11.1% on Longitudinal Braking, both in the *same* Spatial Alignment primitive. These uneven profiles show that success on one task does not carry over to closely related tasks in the same primitive.

Table 5: **Single vs. paired failure rates (%) by dimension and model** (pooled over three generation runs). Failure rate = fraction of instances scored 0; for the single-video columns an instance counts as failed if either of its two videos fails, so these rates are higher than 100 minus the per-video pass rates in Table 2. The Δ column shows paired $-$ single; positive values (red) indicate paired evaluation is harder, negative (blue) indicates single is harder. P shows the largest positive Δ (paired much harder); E shows a negative Δ (correlated drift makes paired easier).

	Adherence			Physics			Environment		
Model	s	p	Δ	s	p	Δ	s	p	Δ
CogVideoX1.5	83.9	80.8	−3.1	65.8	87.9	+22.0	81.7	48.4	−33.3
Wan2.2	80.4	77.5	−2.8	59.8	84.6	+24.9	66.1	37.9	−28.2
HunyuanVideo1.5	82.1	79.0	−3.1	50.6	86.2	+35.6	61.2	35.3	−25.9
Cosmos-Predict2	78.5	74.5	−4.0	56.0	79.8	+23.8	54.8	24.5	−30.3
Seedance 1.5*	70.5	69.9	−0.6	63.2	81.0	+17.8	77.1	53.1	−24.0
Kling 3.0*	67.8	64.4	−3.5	51.4	73.7	+22.3	47.7	23.4	−24.3
Seedance 2.0*	59.2	56.9	−2.3	42.7	68.5	+25.8	38.6	28.7	−9.8
Veo 3.1*	52.4	49.6	−2.7	47.9	57.5	+9.6	48.7	33.2	−15.5
Grok Imagine*	53.2	49.7	−3.5	45.8	57.7	+11.9	45.5	26.0	−19.6
Mean	69.8	66.9	−2.8	53.7	75.2	+21.5	57.9	34.5	−23.4

C.3 Failure-Mode Labeling

For every failing check, Gemini 3.1 Pro read the judge's explanation (text only, no video, nothing re-judged) at temperature 0 and assigned any of the modes in Table 6; labels can overlap, or be "other". The final-state description for O_p mode collapse below labels explanations of already-failed checks; it is not the O_p pass rule, which allows the expected difference at any point in the video. The definitions are the same for all runs and models, and the percentages in Table 4 are shares of failing checks, averaged over the three runs (run-to-run SD at most 2.4 pp). As a reliability check we repeated the labeling with Gemini 2.5 Flash. The two passes agree with $\kappa \geq 0.65$ on 11 of the 17 modes and 0.50–0.58 on six (action omission, action substitution, ghost force, paired-physics mode collapse, environment divergence, divergent hallucination), a moderate level of agreement consistent with the labels' use as a descriptive breakdown. The "other" share is 9–41% depending on the dimension and is left out of Table 4.

C.4 Effect of Outcome Hints in the Generation Prompt

Our main protocol (Section 3.2) deliberately withholds the physical outcome from the generation prompt, so that the model must simulate the consequence of the intervention from x_0 and v rather than retrieve it from textual priors. To test whether explicitly stating the expected outcome improves the generated contrast, we run a controlled ablation on Grok-generated videos in which the generation prompt is augmented with an outcome statement.

We compare two configurations: (1) **Base**, the standard protocol where prompts contain only camera perspective, scene state, and intervened action (no outcome wording), and (2) **+Hint**, where the same prompts are extended with a brief clause describing the intended physical outcome (e.g., *"...causing the gap between the ego car and the leading vehicle to increase"*). The judge configuration and all evaluation procedures are held identical across the two conditions. Both are scored on the same 269 pairs from 17 sub-categories, with Base taken from run 1.

Including the outcome in the generation prompt produces consistent but *modest* gains across all four dimensions (+0.042 on average), with the largest improvements on O_p (+0.059) and E_p (+0.045) and the smallest on P_p (+0.026).

Table 6: **Failure-mode definitions used for labeling**, and κ between the Gemini 2.5 Flash and Gemini 3.1 Pro labeling passes. Definitions are shortened.

Check	Mode	Definition	κ
A single	Action omission	The prompted action is not carried out at all; the scene stays static.	0.58
	Action substitution	A different action is carried out, by the wrong entity or in the wrong direction.	0.58
A paired	No discernible difference	The two videos look the same, or the difference is negligible.	0.79
	Baseline inconsistency	The two videos do not share a starting state or motion baseline.	0.88
P single	Kinematic discontinuity	Teleportation, sudden jumps in speed or position, impossible velocities.	0.80
	Morphing	Objects, arm, gripper or vehicles change shape or identity, or clip and duplicate.	0.87
	Ghost force	Objects move or float with no contact or cause.	0.50
P paired	Mode collapse	Same or near-identical trajectory in both videos; no divergence at the intervention.	0.58
	No interaction	The required contact, grasp, collision or push never happens in one or both videos.	0.65
E single	Object permanence	Objects appear, disappear, duplicate or change identity within a video.	0.72
	Background collapse	The background warps, dissolves or degrades.	0.86
	Camera hallucination	The camera angle, height or framing changes abruptly without cause.	0.93
E paired	Environment divergence	Background, camera or non-target objects differ between the two videos.	0.55
	Divergent hallucination	One video adds objects or textures where the other has none.	0.58
O paired	Mode collapse	Both videos end in the same, indistinguishable final state.	0.85
	Upstream propagation	An earlier failure (no action, no contact, morphing) means the expected event never happens.	0.70
	Reversed outcome ("causal inversion")	The outcome differs but in the wrong direction.	0.87

Table 7: Paired-comparison scores for **Grok** under **Base** (outcome withheld from the generation prompt) vs. **+Hint** (outcome included in generation prompt), both on the same 269 pairs from 17 sub-categories (run 1). The VLM judge is identical in both conditions; only the generation prompt differs. Δ: absolute change.

Dimension	Base	+Hint	Δ
Adherence (A_p)	0.498	0.535	$+0.037$
Physics (P_p)	0.450	0.476	$+0.026$
Environment (E_p)	0.680	0.725	$+0.045$
Outcome (O_p)	0.461	0.520	$+0.059$
Average	0.522	0.564	$+0.042$

The hint does not close the gap. Even when the model is explicitly told what should happen, O_p rises only to 52.0% and P_p only to 47.6%. The modest improvement shows that stating the outcome alone is insufficient to generate the prescribed physical transition reliably.

C.5 Per-Sub-Category Outcome Scores

Table 8 provides the most granular view of model performance: every model's outcome score (O_p) on every sub-category, sorted from easiest (top) to hardest (bottom).

No model is uniformly strong. Even the top-ranked Grok Imagine scores 0% on Overtake Accel and 11% on both Long. Brake and Overtake Brake, yet reaches 98% on Accel Intensity. This spread (from near 0% to near 100% within the same model) shows that aggregate scores should be read together with the per-intervention breakdown.

The bottom of the table is almost entirely blue or white. Sub-categories like Object Mass (hardest, 7.9% average), Overtake Hard (11.4%), Table Friction (11.9%) and Overtake Accel (12.3%) are below 25% for most models. These tasks share a common property: the effect of the variable is either subtle (mass, friction coefficient) or requires relational reasoning across agents (timing). In contrast, the top of the table (Acceleration Intensity, Longitudinal Acceleration) involves visually dramatic scenarios where most closed-source models achieve 50–100%.

Open-source models (right-hand columns) rarely reach the higher tiers. Only 3 of their 72 cells reach the red tier ($\geq$50%): Wan2.2 and Cosmos-Predict2 on Acceleration Intensity, and Wan2.2 on Longitudinal Acceleration. Closed-source models populate more red and orange cells, but even they collapse on the hardest sub-categories.

Table 8: **Per-model O_p (%) on every sub-category** (pooled over three generation runs), sorted by difficulty (easiest top). Cells are color-coded: $\geq$50 , 25–49 , 1–24 , white = 0. n = number of pairs.

Sub-category	n	Grok	Veo 3.1	Seedance 2.0	Seedance 1.5	Kling	Wan	Cosmos-Predict2	CogVideoX	Hunyuan
Accel Intensity	14	98	98	100	79	74	64	64	36	38
Long. Accel	18	93	98	52	57	37	69	19	48	20
Release Coords	20	68	62	50	23	33	17	32	12	22
Deformability	15	47	64	31	38	31	11	29	0	18
Grasp Align	22	42	48	27	18	38	15	41	8	21
Brake Intensity	18	81	63	26	37	19	2	13	9	6
Gripper Timing	22	29	32	42	11	38	23	30	12	15
Long. Brake	18	11	46	43	43	17	20	6	28	11
Push/Yank	20	37	37	23	38	18	10	17	8	13
Scene Layout	15	80	42	29	18	16	4	0	2	0
Road Medium	15	48	76	2	22	11	11	2	4	7
Lateral Avoid	17	57	43	27	8	16	0	2	0	0
Overtake Brake	15	11	24	27	9	13	13	7	18	4
Road Friction	15	43	9	18	11	18	7	2	9	7
Overtake Accel	15	0	11	31	9	13	11	2	22	11
Table Friction	15	33	13	20	11	18	2	2	2	4
Overtake Hard	30	28	40	14	7	4	6	1	1	1
Object Mass	15	20	20	4	7	7	2	7	2	2

C.6 Which Changes Models Miss

Section 4.3 shows where contrastive reasoning is missing and why the gaps are hard to see: they sit next to successes, and a change to the world is often shown without its effect. This section gives the numbers behind both points. Many of the missed changes also share one property: the correct effect comes from something that resists the motion, not from the agent pushing harder. Values are means over the nine models, each averaged over the three runs. Intervals use the scene-cluster bootstrap of Appendix F.3.

Grouping. Table 10 lists the 18 sub-categories by O_p. We mark as *drive* the three whose intervention adds force to the agent's own motion (accelerating, or pushing harder) and as *resist* the eight whose effect comes from something that opposes motion: braking, friction, a dense or deformable material, an obstacle, or the mass of the object. The other seven change where or when the agent acts and

Table 9: **Outcome when both videos follow their own prompts** (%, pooled over the nine models and three runs). Both A_s: share of pairs in which both videos pass A_s. $O_p \mid A_s$: O_p pass rate among those pairs; $O_p \mid A_s, P_s$: the same when both videos also pass P_s. Last column: share of O_p failures that come from pairs in which both videos pass A_s.

	Both A_s	$O_p \mid A_s$	$O_p \mid A_s, P_s$	O_p fails with both A_s
Action primitives	28.6	80.9	95.0	7.6
Spatial Alignment	32.8	81.9	95.7	8.8
Force/Degree	34.7	90.5	96.9	5.3
Temporal Sequencing	19.8	68.1	90.9	7.5
Environment primitives	34.3	45.8	85.9	22.5
Surface Friction	36.7	28.7	75.8	29.9
Material/Medium	29.0	80.6	97.7	7.5
Obstacle Configuration	37.5	34.4	76.5	28.7
All	30.2	69.6	92.8	12.2

belong to neither group; they are hard as well, so the drive/resist contrast applies only to force-related interventions. This grouping is descriptive; two comparisons defined by the benchmark itself show the same pattern: acceleration against braking on the same scenes, and the taxonomy's split into action and environment primitives. (Push/Yank, whose outcome is read from how the pushed object or the surface responds, is the one drive sub-category that scores low, O_p 22.4%.)

Speeding up and slowing down. Longitudinal Acceleration and Longitudinal Braking use the same 18 scenes, and every model passes O_p more often on acceleration (54.7% against 24.9%, a gap of 29.8 pp with 95% CI [23.5, 35.4]; Table 11). The intensity sub-categories show the same ordering for every model (72.2% against 28.4%, a gap of 43.8 pp, 95% CI [36.7, 51.2]). Single-video physics does not separate the two directions (P_s about 85% in all four sub-categories; the accel-minus-brake difference is -1.2 pp, 95% CI [-4.0, 1.4], on the longitudinal pairs and 1.2 pp, [-3.0, 5.3], for intensity), while paired physics does (P_p 45.3% against 20.0%, and 63.2% against 23.5%). Part of the braking failure is that the brake is not carried out as prompted (A_s 47.2% against 67.3% on the longitudinal pairs), which single-video adherence can see; single-video physics cannot. In overtaking, where the timing decides the outcome, both directions stay near the floor. Grok Imagine shows the largest contrast, 92.6% against 11.1% on the longitudinal pairs (gap 81.5 pp, 95% CI [68.5, 92.6]).

The agent's action and the environment. All three environment primitives act through resistance, and they reach O_p = 17.5% against 27.7% for the action primitives (Appendix C.7). The two groups also fail differently. For the action primitives the outcome pass rate is close to the rate at which the intervention is rendered (A_p 30.5%, O_p 27.7%). For the environment primitives the change is rendered more often (A_p 39.5%, partly because 72 of these pairs show it in the first frame), but the outcome follows less than half as often (O_p 17.5%): the model shows the snow or the cone, and the motion does not change with it. When the change is given only in the text, it is rarely rendered at all (A_p 10.7% on the 18 text-only pairs; Table 13).

When both videos follow their prompts. A pair can fail O_p because one of its videos does not follow its own prompt, which a single-video adherence check would also catch, or because both videos follow their prompts and the pair still does not show the expected difference. Table 9 separates the two. Among pairs in which both videos pass A_s, the outcome is correct in 80.9% of action pairs but only 45.8% of environment pairs (gap 35.1 pp, 95% CI [26.1, 43.2]), and in 28.7% and 34.4% for Surface Friction and Obstacle Configuration. Requiring P_s for both videos as well raises these rates, and the gap remains (95.0% against 85.9%, 95% CI of the gap [1.2, 18.0]). For environment interventions, failures that single-video adherence does not reveal are therefore common: 22.5% of their O_p failures come from pairs in which both videos follow their prompts, against 7.6% for action interventions. Rates are pooled over the nine models and three runs; intervals use the scene-cluster bootstrap of Appendix F.3.

Table 10: **Sub-categories grouped by what the intervention changes**, sorted by O_p (%, mean over the nine models and three runs). Drive: the intervention adds force to the agent's motion. Resist: the effect comes from something that opposes motion. Where / when: position or timing.

Sub-category	Domain	Pairs	Group	P_s	O_p
Accel Intensity	AD	14	drive	85.3	72.2
Long. Accel	AD	18	drive	84.7	54.7
Release Coords	Rob	20	where / when	69.4	35.4
Deformability	Rob	15	resist	47.3	29.9
Grasp Align	Rob	22	where / when	61.8	28.8
Brake Intensity	AD	18	resist	84.2	28.4
Gripper Timing	Rob	22	where / when	58.2	25.8
Long. Brake	AD	18	resist	85.9	24.9
Push/Yank	Rob	20	drive	45.3	22.4
Scene Layout	AD	15	resist	55.3	21.2
Road Medium	AD	15	resist	56.6	20.4
Lateral Avoid	AD	17	where / when	30.1	17.0
Overtake Brake	AD	15	where / when	64.3	14.1
Road Friction	AD	15	resist	36.7	13.7
Overtake Accel	AD	15	where / when	60.2	12.3
Table Friction	Rob	15	resist	48.5	11.9
Overtake Hard	AD	30	where / when	84.9	11.4
Object Mass	Rob	15	resist	39.5	7.9

Table 11: **Accelerating and braking in driving.** O_p (%) per model on the six driving sub-categories in which the ego car accelerates or brakes (pooled over three generation runs), and means over the nine models of O_p, A_s, P_s and P_p. Long.: close or open the gap to a lead vehicle (Spatial Alignment; both use the same 18 scenes); Intensity: how hard (Force/Degree); Overtake: timing in an overtake (Temporal Sequencing). * closed-source.

	Accelerate			Brake		
Model	Long.	Intensity	Overtake	Long.	Intensity	Overtake
CogVideoX1.5	48.1	35.7	22.2	27.8	9.3	17.8
Wan2.2	68.5	64.3	11.1	20.4	1.9	13.3
HunyuanVideo1.5	20.4	38.1	11.1	11.1	5.6	4.4
Cosmos-Predict2	18.5	64.3	2.2	5.6	13.0	6.7
Seedance 1.5*	57.4	78.6	8.9	42.6	37.0	8.9
Kling 3.0*	37.0	73.8	13.3	16.7	18.5	13.3
Seedance 2.0*	51.9	100.0	31.1	42.6	25.9	26.7
Veo 3.1*	98.1	97.6	11.1	46.3	63.0	24.4
Grok Imagine*	92.6	97.6	0.0	11.1	81.5	11.1
Mean O_p	54.7	72.2	12.3	24.9	28.4	14.1
Mean A_s	67.3	75.1	36.8	47.2	56.1	36.5
Mean P_s	84.7	85.3	60.2	85.9	84.2	64.3
Mean P_p	45.3	63.2	12.3	20.0	23.5	11.4

C.7 Action and Environment Interventions

The comparison in Section 4.3 uses the taxonomy's own split, fixed before any model was run: interventions on the agent's action ($\mathcal{D}_{\text{act}}$: Spatial Alignment, Force/Degree, Temporal Sequencing) and interventions on a property of the world ($\mathcal{D}_{\text{env}}$: Surface Friction, Material/Medium, Obstacle Configuration). Table 12 shows the gap in both domains; it also holds for eight of the nine models (Grok Imagine ties).

Table 12: O_p **(%) for action and environment interventions**, mean over the nine models and three runs. The gap is 10.2 pp (95% cluster-bootstrap CI [6.2, 14.4]).

	All pairs	Driving	Manipulation
Action primitives	27.7	27.5	28.0
Environment primitives	17.5	18.5	16.5

One reading of why environment interventions score lower is that an action intervention changes the agent's own motion, which is a large visible difference, whereas an environment intervention acts through a property whose effect on the vehicle or the object is subtler, even when the property itself is shown (snow, mud or a cone in the edited frame, or a heavy block described as carrying a "100 KG" label).

In the 72 environment pairs with an edited frame the two videos start from different images, and E_p is lower on these pairs (51.1% against 65.0%; Table 13). The 18 text-only pairs score lower on O_p than the edited ones, but the subsets differ in scenes and interventions, so this comparison does not isolate the effect of image editing.

Table 13: **Environment pairs with an edited frame and with text only** (mean over the nine models and three runs, %).

	Pairs	A_p	P_p	E_p	O_p
Edited frame (two images)	72	46.8	17.5	51.1	19.0
Text only (one image)	18	10.7	10.9	65.0	11.3

D Benchmark Construction

D.1 Benchmark Statistics

The benchmark comprises 319 test pairs spanning six physical primitives across two embodied domains: Autonomous Driving (AD, 11 sub-categories) and Robotic Arm Manipulation (Rob, 7 sub-categories), for 18 sub-categories in total. Each pair is run on all 9 video generation models (4 open-source, 5 closed-source) with two generations per pair. Some pairs share a baseline video: in the longitudinal sub-categories (three prompts per scene) and the overtaking sub-categories (five prompts per scene), several pairs reuse the same baseline generation. Counting shared videos once, a complete run requires 575 unique videos per model, so 5,175 videos for the nine models. The three-run design therefore comprises 15,525 unique model–run video slots before unavailable outputs and evaluation exclusions (Appendix F.2); these design counts are distinct from the denominators used for reported scores. Table 15 summarizes the composition.

D.2 Taxonomy Construction Protocol

This section provides the full construction trace for the six-primitive taxonomy introduced in Section 3.1, following the thematic-analysis framework (Braun & Clarke, 2006) adapted for the construction of a discrete space of physical variables.

Stage 1: Random Video Sampling. We drew 1,000 candidate videos by simple random sampling: 500 driving clips from nuScenes (Caesar et al., 2020) and 500 manipulation episodes from

Table 14: **The Unified Physical-Variable Taxonomy across Embodied Domains.** Six physical primitives organized into two groups (Environment and Action), three per group, each instantiated in both Autonomous Driving and Robotic Manipulation. Citations indicate prior physical-reasoning benchmarks grounding each primitive.

Group	Physical Primitive	Autonomous Driving	Robotic Manipulation
$\mathcal{D}_{\text{env}}$ *(Environment)*	Surface Friction (Patel et al., 2022; Tung et al., 2023)	Dry asphalt vs. snowy / icy road	Smooth (glass) vs. grippy (rubber mat) surface
	Material/Medium (Gao et al., 2024a; Zheng et al., 2024)	Asphalt vs. muddy road	Rigid object (wood) vs. deformable object (sponge)
	Obstacle Configuration (Bakhtin et al., 2019)	Clear path vs. static obstacle ahead (cone, pothole)	Light, movable target vs. target too heavy to push
$\mathcal{D}_{\text{act}}$ *(Action)*	Spatial Alignment (Girdhar & Ramanan, 2019)	Gap to a lead vehicle (constant speed vs. accelerating or braking); lateral lane offset (safe merge vs. collision)	Gripper closure position (grasp success vs. miss)
	Force/Degree (Ates et al., 2022; Liu et al., 2024a)	Light vs. hard braking; light vs. heavy acceleration	Small vs. large rotation angle; gentle vs. forceful push
	Temporal Sequencing (Girdhar & Ramanan, 2019; Grunde-McLaughlin et al., 2021)	Speed / timing relative to leading car or merge point	Align gripper before closing vs. premature closure

Table 15: **Dataset composition.** Number of pairs and planned model–pair instances (pairs $\times$ 9 models) per sub-category and per run, before exclusions. AD = Autonomous Driving; Rob = Robotic Manipulation. Scene 14 of Object Mass is scored under Material/Medium, so the primitive-level tables count 29 pairs for Obstacle Configuration and 31 for Material/Medium.

Primitive	Sub-category	Domain	Pairs	Instances
Spatial Alignment	Long. Accel	AD	18	162
	Long. Brake	AD	18	162
	Lateral Avoid	AD	17	153
	Release Coords	Rob	20	180
	Grasp Align	Rob	22	198
Force/Degree	Brake Intensity	AD	18	162
	Accel Intensity	AD	14	126
	Push/Yank	Rob	20	180
Temporal Sequencing	Overtake Hard	AD	30	270
	Overtake Accel	AD	15	135
	Overtake Brake	AD	15	135
	Gripper Timing	Rob	22	198
Surface Friction	Road Friction	AD	15	135
	Table Friction	Rob	15	135
Material/Medium	Road Medium	AD	15	135
	Deformability	Rob	15	135
Obstacle Configuration	Scene Layout	AD	15	135
	Object Mass	Rob	15	135
Total			**319**	**2871**

DROID (Khazatsky et al., 2024). Sampling used a fixed seed for reproducibility. We did not pre-stratify by scene type or task category, so the pool reflects the diversity of the two source datasets.

Stage 2: Independent Open Coding. Two domain experts (one in autonomous driving, one in manipulation) independently reviewed their respective 500-video subset. For each clip, they recorded all physical variables whose perturbation would plausibly change the visible outcome. The AD expert proposed 18 candidate variables; the manipulation expert proposed 14, yielding 32 initial candidates across the two domains.

Stage 3: Joint Adjudication and Merging. The two experts then convened to (a) apply the three admission criteria stated in Section 3.1—physically fundamental, embodiment-shared, operationally isolable—and (b) merge candidates that shared the same underlying physical mechanism. Two candidates were merged iff they (i) acted through the same physical mechanism on the outcome, and (ii) admitted a unified contrastive-intervention template. During adjudication, some candidates were rejected outright for failing an admission criterion, and others were absorbed into a broader sub-category as redundant within-domain duplicates. The remaining candidates were then organized into six unified primitives.

A representative rejection is *time of day*, along with the closely related *ambient lighting*. We excluded both for failing the *physically fundamental* criterion, since they are perceptual variables that change how the scene is rendered rather than physical mechanisms governing how force translates into motion, and they cannot be cleanly isolated from co-varying properties such as shadows and contrast.

Representative within-domain merges include:

- In AD, *left-lane shift* and *right-lane shift* were merged into the direction-agnostic sub-category *Lateral lane offset*, since both act through the same lateral-misalignment mechanism with sign-symmetric outcomes.
- In AD, four overtaking-related variants, *side passing*, *overtake-then-merge*, *decelerate to yield*, and *decelerate-then-merge*, were collapsed into the unified sub-category *Timing of overtaking vs. yielding*, since all four reduce to controlling *when* the ego vehicle initiates its longitudinal action relative to a parallel-traveling neighbor; the optional follow-up lane change does not introduce a new physical variable.
- In manipulation, *top-down grasp* and *side grasp* were merged into *3D grasp alignment*, since both act through the same gripper–object spatial-alignment mechanism and differ only in approach direction.
- In manipulation, *object weight* and *object density* were merged into *Object mass*, since the variable that ultimately governs the contact dynamics is the integrated mass rather than its decomposition into volume and density.

Cross-domain unification then proceeded at the primitive level. For instance, braking intensity in AD and gripper push intensity in manipulation were placed under the same *Force/Degree* primitive despite acting in different embodiments, because they share the same mechanism: a continuous scalar on the action side that monotonically scales the outcome.

Stage 4: Reliability via Unanimous Consensus. Because the two experts independently coded *disjoint* domain-specific subsets at Stage 2, we enforce a strict validity criterion at Stage 3: every retained sub-category and final primitive grouping was the product of *unanimous agreement* between the two domain experts. Disagreements during adjudication were resolved by joint discussion until consensus, and no category was admitted by majority vote or external arbitration.

At the implementation stage of benchmark construction (Section 3.2), some sub-categories were further split for convenience of test-instance authoring (e.g., *Longitudinal relative distance control* was instantiated separately as longitudinal acceleration and longitudinal braking variants), yielding the 18 sub-categories used to organize the 319 test pairs in Table 15.

D.3 x_0 Selection Protocol

This section provides the construction trace for the visual anchors x_0 of the 319 pairs. The protocol implements the two requirements of Section 3.2, that x_0 lie at the branching point and that the clip

carry the physical affordances the target sub-category needs, through a three-stage pipeline that combines automated screening with human consensus.

Stage 1: VLM-based candidate selection. For each of the 18 sub-categories in Table 15, we drafted a sub-category-specific eligibility specification listing (i) the physical affordances required for the target intervention to be testable (e.g., for *Accel Intensity*: ego vehicle stationary or moving slowly with a clear lane ahead; for *Grasp Alignment*: gripper open, target object unoccluded and within reachable workspace), (ii) the visibility requirements (agent and target fully visible, no heavy occlusion or motion blur, ego vehicle not parked, pre-contact gripper state in manipulation), and (iii) the definition of action onset for the intervention. We then prompted Gemini 3.1 Pro (Google DeepMind, 2026) to scan candidate clips from nuScenes (Caesar et al., 2020) and DROID (Khazatsky et al., 2024) against these specifications. For each eligible clip, the VLM additionally returned a candidate timestamp identifying the frame immediately preceding action onset, which we extract as the candidate x_0. Ineligible clips (e.g., missing affordances, insufficient visibility, or with the action already underway) were filtered out and not forwarded to human review.

Stage 2: Independent human review. Two researchers then independently reviewed every Stage 1 candidate (clip plus VLM-suggested x_0) and rendered a binary keep/reject decision. A candidate was rejected if any of the following held: (i) *visibility*: occlusion, motion blur, or lighting obscures the target agent or object; (ii) *ambiguous spatial relationships*: the frame does not unambiguously specify the geometry between agent and target (e.g., lateral offset between two vehicles, gripper–object alignment), so $\mathcal{V}_+$ and $\mathcal{V}_-$ would not share a well-defined initial geometry; (iii) *state-fixation violation*: the candidate frame is mid-action rather than pre-action, so v has already begun to materialize; or (iv) *affordance mismatch*: the frame does not actually support the intended intervention despite Stage 1 admission. Visibility and ambiguous spatial relationships accounted for most rejections.

Stage 3: Unanimous consensus. Only candidates retained by *both* reviewers were admitted to the benchmark; candidates kept by only one reviewer were discarded without further adjudication. This rule trades quantity for cleanliness: a frame whose admissibility is contested is precisely the kind whose initial state cannot be guaranteed fixed across $\mathcal{V}_+$ and $\mathcal{V}_-$, and admitting it would weaken the state fixation that paired evaluation relies on.

The three screening gates. Stages 1 and 2 apply one written rule: a candidate clip is kept for a primitive only if it passes three gates.

1. x_0 **admits both branches.** The main text asks for "the physical affordances that make a target primitive testable." We apply this as a symmetric rule: both the p_+ and the p_- continuation must be plausible futures of the same scene. If x_0 already settles the variable, one branch contradicts the frame. For example, a dry-versus-wet friction pair rejects a frame that already shows rain, since only the wet branch is then plausible.
2. **The intervention's effect is visible.** The main text prefers that any divergence stays visible against an uncluttered scene. We used this as a hard gate. The paired judgments compare $\mathcal{V}_+$ and $\mathcal{V}_-$ against the divergence the intervention predicts, and if that divergence is tiny or buried in clutter, the item measures rendering limits and not contrastive reasoning. For example, a mass contrast needs a push strong enough that mass visibly matters, not a near-zero nudge.
3. **No scene-side confounder.** The single-variable rule is stated for the prompts, and we applied it to the scene too: no other element's continuation may depend on the target variable. For example, a friction pair excludes scenes with a lead vehicle whose braking would also change with the surface, because that varies two variables at once.

Per-primitive checklists. The gates were turned into a must-have / must-not-have list for each primitive (Table 16).

Environmental isolation via generative augmentation. Some environment interventions cannot be expressed in text alone. A real frame of a sunny road contradicts a prompt that says the road is icy, and a prompt about an obstacle cannot add the obstacle to a frame that has none. For these pairs one prompt is paired with the original frame and the other with an edited copy of the same frame, made with Gemini 3 Pro Image (Google, 2025a), in which only the target environmental condition is changed (for example rendering snow, or adding a traffic cone) and everything else (scene layout, objects and their positions, agents, camera viewpoint) is kept unchanged. This is the case for 72 of the 319 pairs (22.6%), all of them in the six environment sub-categories. The other 247 pairs,

Table 16: **Per-primitive screening checklists.** Written from the three gates.

Primitive	Must have	Must not have
Spatial Alignment	at least one other vehicle or object to interact with; ego moving at steady speed (driving) or an open gripper near an unoccluded target (manipulation); lane or workspace clearly visible	a lead vehicle so close that there is no room to act; heavy occlusion; a gripper that is already grasping
Force/Degree	ego stationary or at steady speed on an open road with room to speed up or slow down, or a clearly visible target for a push; the vehicle's rear or front visible so stopping distance can be judged	a lead vehicle that limits the action (a light and a hard brake would then look alike); clutter around the pushed object
Temporal Sequencing	a parallel vehicle or forward target with clearly visible relative speed and distance, or an open gripper that has not yet closed	a merge or lane-change opportunity that has already passed; an object that is already grasped
Surface Friction	ego in motion on a homogeneous surface with room to act	pre-existing wet or snow; brake input visible at x_0; multi-material transitions
Material/Medium	a single, clearly visible road material, or a known rigid versus deformable object whose deformation can be seen	mixed road surfaces (mud on asphalt) that make the baseline unclear; occluded objects
Obstacle Configuration	a clear path as baseline with ego going straight, or an empty workspace; for mass, a target the arm is pushing with enough force that the result depends on mass	an obstacle that is already in the path (it would fix the baseline)

including 18 environment pairs whose intervention can be written in words alone, use one and the same frame for both prompts (Appendix D.5).

The edited pairs remain directly comparable because the two frames differ only in the variable under intervention (Appendix D.5).

Per-sub-category sampling band. We targeted about 15 to 25 scenes per sub-category and added Stage 1 candidates when a sub-category fell short after Stage 3 attrition (final counts in Table 15).

D.4 Prompt Authoring Protocol

All 319 contrastive pairs were authored under a three-stage human consensus protocol that follows the construction requirements specified in Section 3.2: state fixation, single-variable change, and outcome withholding.

Stage 1: Independent drafting. Two researchers independently drafted (p_+, p_-) pairs for each candidate clip, applying the three-part template introduced in Section 3.2 (camera perspective, initial scene state, agent action) and intervening only within the span that describes the target variable.

Stage 2: Cross-review. A third researcher reviewed every drafted pair against the Section 3.2 criteria, flagging any pair whose action description was inaccurate, inconsistent with the visible content of x_0, or potentially leaking the outcome.

Stage 3: Unanimous adjudication. All three researchers jointly discussed the flagged pairs, and only pairs receiving unanimous approval were retained. Approximately 37% of initial drafts were rejected at this stage. The two dominant rejection reasons were (i) the action description did not faithfully match the affordances visible in x_0, and (ii) the wording was insufficiently specific to make the intended intervention unambiguous. Both reasons weaken state fixation, since if the prompt does not match what x_0 shows, $\mathcal{V}_+$ and $\mathcal{V}_-$ no longer originate from the same effective initial condition. We therefore treated these as hard rejections rather than candidates for revision.

Surviving pairs were then normalized to a uniform stylistic format (third-person present tense, fixed action-span opener) so that format unification did not reintroduce cross-pair content variance. As

a final independent check on the single-variable-change requirement of Section 3.2, every retained pair was passed through Gemini 3.1 Pro acting as an automated prompt-pair evaluator. For each pair, Gemini received p_+, p_-, and the name of the target variable v (for example, "braking intensity" or "surface friction"), and was instructed to judge whether the two prompts describe the same underlying scene, camera setup, objects, and initial conditions, and differ in exactly one aspect, namely the description of v. The instruction explicitly admitted minor wording variation that preserves meaning (synonyms, paraphrasing, equivalent grammatical constructions) as acceptable, since such variation does not introduce additional physical factors capable of affecting the generated video; only differences that would constitute an additional controlled variable were treated as violations. Gemini returned a binary judgment, together with, for any flagged pair, a list of the additional aspects in which the two prompts diverged. Flagged pairs were manually re-reviewed and either revised or rejected before being added to the released benchmark.

The three-way human consensus is the primary check of pair validity; Gemini adds a second one.

D.5 Initial Frames and Object Mass

Initial frames. We audited the image manifest of all 319 pairs. All 229 action pairs use one and the same file for both prompts (byte-identical). Of the 90 environment pairs (six sub-categories, 15 each), 72 use two files, a baseline and an edited copy, and 18 use one file because their intervention can be written in words. Every edited frame was checked by a person against its original and its edit instruction, and only the intended change was found; the side-by-side pictures are released.

Object Mass. Action intensity (e.g., a harder push) belongs to Force/Degree; target-object mass belongs to Obstacle Configuration. One of the 15 Object Mass scenes (a loose against a glued pile of hats) tests material cohesion, so we score it under Material/Medium.

E VLM Judge

We evaluate each contrastive video pair using a Vision–Language Model (VLM) judge with a structured system prompt. The prompt follows one template, instantiated per sub-category. Below we present the generalized template and its instantiation for one sub-category (Category 1: Ego Following Speed Change, i.e., Long. Accel). In the prompt, Video A and Video B denote $\mathcal{V}_+$ and $\mathcal{V}_-$.

E.1 Generalized Template Structure

Each evaluation prompt consists of five sections:

1. **Role Assignment & Task Framing.** The VLM is instructed to act as an *Autonomous Driving Video or Robotic Arm Evaluator and Causal Inference Specialist*, comparing two dashcam videos generated from text prompts that differ in a single causal variable.
2. **Scenario Context.** A category-specific block specifying: (a) the *target entity* undergoing the causal intervention, (b) *context entities* that must remain invariant, (c) the *intervention* applied in Video A vs. Video B, and (d) *causal constraints* dictating which entities may and may not change.
3. **APEO Rubric.** Four evaluation dimensions, each with single-video and/or paired checks, strict binary (0/1) scoring, failure-trigger lists, and chain-of-thought (CoT) reasoning examples.
4. **Scoring Logic.** A strict pass/fail protocol: Score 1 requires flawless execution; *any* single violation (ghosting, semantic bleed, missing action, background instability) triggers Score 0.
5. **Output Schema.** A fixed JSON schema with CoT observations and a binary score per check.

E.2 APEO Evaluation Dimensions

Each dimension corresponds to a downstream use of a video world model. *Adherence (A)* matters for VLA policy training (Wang et al., 2025b; Kim et al., 2024; Zitkovich et al., 2023) and action-conditioned world models (Yang et al., 2023; Bruce et al., 2024): if $\mathcal{V}_-$ silently performs a different action, the policy learns a wrong action–outcome map. *Physics (P)* matters for inverse dynamics learning (Baker et al., 2022) and latent dynamics modeling (Hafner et al., 2023): physically impossible

intermediate frames corrupt the dynamics signal even when endpoints look correct (Kang et al., 2025). *Environment (E)* matters for synthetic data augmentation (Wang et al., 2023; Hu et al., 2023): hallucinated background changes induce shortcut learning (Geirhos et al., 2020). *Outcome (O)* matters for model-based planning and counterfactual simulation (Hafner et al., 2023; Yang et al., 2023): their value depends on rolling out distinct futures under distinct actions.

Dimension Definitions (Generalized)

[A] Intervention Adherence (The Cause) — Does each video correctly depict its assigned causal intervention without spatial misalignment, temporal omission, or target confusion?

- *Single-Video*: Score each video independently on whether the instructed action is visible and correctly applied to the target entity.
- *Paired*: Is there a visually discernible difference between Video A and Video B that reflects the intended causal manipulation?
- *Failure triggers*: Action Omission, Target Misalignment, Temporal Lag, Pseudo-Action Illusion, Lateral Redirection.

[P] Physics — Does the causal transition follow physically plausible rigid-body kinematics and non-holonomic vehicle constraints?

- *Single-Video*: Check each video for kinematic violations, topology/ghosting artifacts, non-holonomic failures, morphing, directional inversion, and oscillatory motion.
- *Paired (Trajectory Bifurcation Alignment)*: The initial segment before the intervention must show identical states; divergence should begin only at the intervention onset.

[O] Counterfactual State Divergence (The Effect) — Does the causally expected outcome manifest at any point during the video?

- *Paired only*: Assess whether the expected state difference (e.g., gap closure, lane position change, object appearance/disappearance) emerges at any point—not restricted to the final frame.
- *Failure triggers*: Mode Collapse (superficial action with no effect), Causal Inversion (opposite outcome), No Delta (identical states throughout).

[E] Contextual & Embodiment Consistency (Environment / Ceteris Paribus) — Do non-causal background variables maintain feature constancy?

- *Single-Video (Absolute Canvas Stability)*: Lane markings, buildings, road surface, bystander vehicle identities, and camera angle must remain stable throughout each video.
- *Paired (Counterfactual Canvas Consistency)*: Masking out the target entity's changed behavior, the baseline road environment and bystander behaviors must be consistent across Video A and Video B.
- *Failure triggers*: Background Collapse, Object Permanence Failure, Camera Hallucination, Phantom Road Geometry, Off-Road Excursion.

E.3 Output JSON Schema

The VLM judge is required to output a single JSON object with the following structure. Each block contains a free-text `observation` field (CoT reasoning) and binary score(s).

Required Output Schema

```
{
  "A_Adherence_Single":   { "observation": "<CoT>",
                            "score_video_A": 0|1,
                            "score_video_B": 0|1 },
  "A_Adherence_Paired":   { "observation": "<CoT>",
                            "score": 0|1 },
  "P_Physics_Single":     { "observation": "<CoT>",
                            "score_video_A": 0|1,
                            "score_video_B": 0|1 },
  "P_Physics_Paired":     { "observation": "<CoT>",
                            "score": 0|1 },
```

```
  "O_Outcome_Paired":    { "observation": "<CoT>",
                           "score": 0|1 },
  "E_Environment_Single":{ "observation": "<CoT>",
                           "score_video_A": 0|1,
                           "score_video_B": 0|1 },
  "E_Environment_Paired":{ "observation": "<CoT>",
                           "score": 0|1 }
}
```

E.4 Category-Specific Instantiation

For each sub-category, the following placeholders in the template are filled in:

Template Placeholders

Scenario Context:
- `[TARGET_ENTITY]` — The entity undergoing the causal intervention (e.g., ego car, traffic light, pedestrian).
- `[CONTEXT_ENTITIES]` — Entities that must remain invariant across both videos (e.g., front car, bystander vehicles, road infrastructure).
- `[INTERVENTION_A]` — Description of the baseline/control behavior shown in Video A.
- `[INTERVENTION_B]` — Description of the causal manipulation shown in Video B.
- `[CAUSAL_CONSTRAINT]` — Which entities may change and which must not.
- `[SCENARIO_VARIANTS]` — Notes on visual variants across samples within this category (e.g., different vehicle types) that do not affect evaluation logic.

Category-Specific Failure Triggers (per dimension):
- `[A_FAILURE_TRIGGERS]` — Adherence failures specific to this category's intervention type.
- `[P_FAILURE_TRIGGERS]` — Physics/dynamics failures relevant to the physical motion involved.
- `[O_FAILURE_TRIGGERS]` — Outcome failures describing what "no causal effect" looks like for this category.
- `[E_FAILURE_TRIGGERS]` — Environment failures specific to the scene elements present.

CoT Examples:
- `[COT_EXAMPLES]` — 2–4 chain-of-thought reasoning examples showing Pass and Fail cases for the most common ambiguities in this category.

E.5 Instantiation Example: Category 1 (Ego Following Speed Change)

For Category 1, the placeholders are filled in as follows. The illustrative wording below is corrected for clarity; reported scores were obtained with the original executed prompts.

Scenario Context — Category 1

[TARGET_ENTITY]: The ego car (camera vehicle).

[CONTEXT_ENTITIES]: A front car driving ahead in the same lane (may be a taxi, bus, sedan, van, or truck). One or more bystander vehicles in adjacent lanes or parked on the roadside. Bystander vehicles *must remain unchanged* across both videos.

[INTERVENTION_A]: The ego car maintains its current speed, keeping a constant following distance from the front car throughout the video.

[INTERVENTION_B]: The ego car suddenly accelerates hard while remaining in its lane, closing the gap to the front car.

[CAUSAL_CONSTRAINT]: The front car and all bystander vehicles must continue at their original speeds in both videos. Only the ego car's speed behavior changes.

[SCENARIO_VARIANTS]: The front car type varies across samples (yellow taxi, orange bus, white bus, sedan, van). This does not affect evaluation logic.

`[A_FAILURE_TRIGGERS]` — Category 1

- *Action Omission*: In Video A, the gap shows clearly obvious and substantial deviation from constant; in Video B, no visible acceleration and gap does not decrease.

- *Target Misalignment*: The front car slows down instead of the ego car accelerating, creating an illusion of gap closure.
- *Temporal Lag*: In Video B, the ego car waits until the final frames to begin accelerating.
- *Pseudo-Acceleration Illusion*: Apparent gap closure caused entirely by road curvature or camera zoom, not actual speed change.
- *Lateral Redirection*: The ego car changes lane to gain speed rather than accelerating longitudinally.
- *Scene-Flow Disambiguation*: When gap closure could be explained by front-car deceleration, use independent scene-flow cues (rate of road markings, ground texture, roadside objects passing the camera) as ego-car velocity proxy. No increase in scene-flow rate means the ego car did not accelerate.

[P_FAILURE_TRIGGERS] — Category 1

- *Kinematic Violation*: Instantaneous speed jump from current speed to high speed in a single frame (no ramp-up).
- *Topology/Ghosting*: Ego car's front bumper phases through the rear of the front car without collision response.
- *Non-Holonomic Failure*: Ego car slides laterally (drifts sideways) during acceleration.
- *Morphing*: Ego car or front car visually distorts, stretches, or changes size during acceleration.
- *Directional Inversion*: Any vehicle travels opposite to its established heading.
- *Oscillatory Motion*: Repetitive forward–reverse–forward motion (generation artifact).

[O_FAILURE_TRIGGERS] — Category 1

- *Mode Collapse*: Ego car in Video B shows visual acceleration cues (e.g., increased scene-flow rate) but the gap never meaningfully closes at any point.
- *Causal Inversion*: Gap in Video B is larger than in Video A at all observable points.
- *No Delta*: Front car appears the same angular size throughout both videos.

[E_FAILURE_TRIGGERS] — Category 1

- *Background Collapse*: Lane markings change count/style; buildings liquefy; road surface changes texture.
- *Object Permanence Failure*: Front car transforms identity (taxi becomes bus); bystander vehicles change lane/speed.
- *Camera Hallucination*: Dashcam angle or height changes abruptly mid-video.
- *Phantom Road Geometry*: New lanes, intersections, or infrastructure materialise mid-video.
- *Off-Road Excursion*: Any vehicle leaves paved road surface.

[COT_EXAMPLES] — Category 1 (selected)

[A] Fail Example:

```
"observation": "In Video B, the front car (orange bus) appeared to
grow larger, but this was because the bus itself slowed down -- the
ego car's scene-flow cues did not increase. The acceleration was
applied to the wrong entity." → score_video_B: 0
```

[P] Fail Example:

```
"observation": "In Video B, the ego car executed a lane change into
an adjacent lane to accelerate, rather than accelerating within
its own lane. Changing lanes violates the single-lane following
constraint." → score_video_B: 0
```

[O] Fail Example:

```
"observation": "Throughout Video B, the front car maintained the same
angular size as in Video A. At no point did the spatial gap noticeably
close despite the ego car appearing to accelerate. The speed change
produced no closing effect." → score: 0
```

[E] Fail Example:

```
"observation": "New road geometry spontaneously appeared mid-video
-- additional lanes materialised where there was previously grass.
Phantom environment generation undermines the shared starting-state
assumption." → score_video_B: 0
```

Table 17: **Human–judge agreement on the ten APEO checks** (421 tuples, run 1). Pass rate = share of tuples marked as passing.

Check	Human pass	Judge pass	Agreement	κ
A, single, video A	72.9	73.6	83.6	0.58
A, single, video B	45.4	42.8	81.2	0.62
A, paired	42.5	40.6	84.3	0.68
P, single, video A	71.7	74.1	86.2	0.66
P, single, video B	50.4	51.8	77.2	0.54
P, paired	32.5	26.1	84.3	0.63
O, paired	34.9	29.0	84.3	0.63
E, single, video A	69.8	69.4	82.9	0.59
E, single, video B	51.5	49.4	79.6	0.59
E, paired	65.3	62.7	81.7	0.60
Overall	53.7	52.0	82.5	0.65

E.6 How the Judge Is Run

Model and call. The judge is Gemini 3.1 Pro (the preview model identifier `gemini-3.1-pro-preview`, through Vertex AI) at temperature 0. For each pair, one request carries the category rubric as the system instruction, video A and video B as two native mp4 attachments (in that order), and a text message with the two generation prompts. The response contains per-video A/P/E scores and paired A/P/E/O scores.

The videos have no audio, so the judge sees the picture only.

What the judge is told about the outcome. The category rubric describes the effect the intervention should have and what a missing or reversed effect looks like (the `[O_FAILURE_TRIGGERS]` blocks above). For the robotic categories the judge also receives a one-sentence description of the expected observable difference for the pair, taken from the prompt file, and is told to use it only for the O_p check. O_p therefore checks whether the described difference appears in the pair; the judge does not have to infer the expected outcome on its own.

Frozen output. For frozen outputs, the prompt template can include the rule: *a video in which nothing changes over time (a frozen frame or an essentially static scene) executes no action and does not count as a stable scene, so it fails the Adherence and Environment checks*.

E.7 Judge Validation

Human labels. Three annotators labeled 421 (model, scene) tuples from run 1, spanning both model tiers and all six primitives, on all ten APEO checks without seeing the judge's answers. Labels were combined by majority vote, with disagreements adjudicated. Inter-annotator agreement (84.1%, κ = 0.67) uses the tuples with at least two annotators. Pass rates stay between 26% and 74% on every check, so κ (Cohen's, against the human label) is not distorted by rare labels.

Agreement with humans. Pooled over the ten checks, κ is 0.65 (82.5% agreement), on par with the 0.67 between annotators. It is 0.54–0.68 on the ten checks, with the paired checks (0.60–0.68) at least as reliable as the single-video ones (0.54–0.66) (Table 17), and 0.59–0.70 on the six primitives (Table 18). On the primitives that might be hard to see, the judge agrees with the human label about as often as annotators agree with each other (Table 18): Surface Friction 84.0% against 84.0%, Material/Medium 83.0% against 84.0%, Obstacle Configuration 80.0% against 77.0%. Temporal Sequencing is the exception (79.6% against 86.8%, κ = 0.59). Across (model, sub-category) groups the judge's pAvg tracks the human pAvg with Pearson r = 0.79 and Spearman ρ = 0.70. The judge is stricter than the humans on all four paired checks, most on paired physics and outcome (pass rates 26.1% against 32.5% and 29.0% against 34.9%), indicating lower judge pass rates in this annotated subset.

Table 18: **Human–judge agreement per primitive** (same 421 tuples). Last column: agreement between annotators on the tuples they both labeled.

Primitive	Tuples	Agreement (%)	κ	Human–human (%)
Spatial Alignment	141	82.6	0.65	85.8
Force/Degree	75	85.1	0.70	83.8
Temporal Sequencing	55	79.6	0.59	86.8
Surface Friction	50	84.0	0.68	84.0
Material/Medium	50	83.0	0.66	84.0
Obstacle Configuration	50	80.0	0.60	77.0
Overall	421	82.5	0.65	84.1

Table 19: **Agreement with the human label, Gemini 3.1 Pro against GPT-5** (κ; 421 tuples, run 1).

	Gemini 3.1 Pro	GPT-5
Spatial Alignment	0.65	0.37
Force/Degree	0.70	0.49
Temporal Sequencing	0.59	0.37
Surface Friction	0.68	0.37
Material/Medium	0.66	0.44
Obstacle Configuration	0.60	0.34
Overall	0.65	0.40

A second judge. GPT-5 scored the benchmark (run 1) with the same rubric and inputs, including every human-labeled tuple; the model comparison uses the 289 pairs per model that both judges scored. It agrees with the human label 70.3% of the time (κ = 0.40) against 82.5% (κ = 0.65) for Gemini; the difference in κ is +0.25 (95% scene-cluster bootstrap interval [0.21, 0.30]), and Gemini is closer to the humans on all six primitives (Table 19). We use Gemini as the main judge because a smaller GPT-5 trial on a labeled subset tracked the humans less closely. The two judges give similar rankings (Table 20; 289 pairs per model): the same three models lead, Kling is fourth and CogVideoX ninth under both, the rankings correlate with Spearman ρ = 0.87, and no model's pAvg differs by more than 8.2 pp. The best model is Grok under Gemini and Veo under GPT-5, so we only claim a joint top tier.

Vendor bias. Gemini ranks Grok Imagine (xAI) first and Veo 3.1 (Google) second, while GPT-5 puts Veo first; Grok is the model Gemini rates highest relative to GPT-5 (-8.2 pp), the opposite of home-team favoritism. The judge sees only the videos and prompts, not the model names, and

Table 20: p**Avg (%) under two judges**, run 1, 289 pairs per model. Ranks are per judge.

	Gemini		GPT-5		
Model	pAvg	Rank	pAvg	Rank	GPT-5 − Gemini
Grok Imagine*	53.0	1	44.8	3	-8.2
Veo 3.1*	50.5	2	48.1	1	-2.4
Seedance 2.0*	44.3	3	45.3	2	+1.0
Kling 3.0*	40.1	4	43.1	4	+2.9
Cosmos-Predict2-2B	33.6	5	32.5	7	-1.0
Seedance 1.5*	33.2	6	38.1	5	+4.8
Wan2.2-5B	28.5	7	30.9	8	+2.4
HunyuanVideo1.5-8.3B	27.2	8	32.6	6	+5.4
CogVideoX1.5-5B	22.3	9	25.9	9	+3.6

Table 21: **Four O_p checks from run 1:** two where the judge agrees with the human label and two where it does not. H / J: human / judge label (1 = pass, 0 = fail).

Case	Pair differs in	H / J	Judge's explanation (excerpt)
Agree, fail HunyuanVideo1.5 Force/Degree	Gentle tap on the gas pedal vs. flooring it	0 / 0	"[. . .] Video B does not cover significantly more ground or move dramatically faster than Video A. The expected counterfactual distance divergence between a gentle tap and full throttle is not present."
Agree, pass Grok Imagine Obstacle Configuration	Light plastic bowl vs. a block marked "100 KG", pushed by the arm	1 / 1	"In Video A, the lightweight plastic bowl is effortlessly sent sliding off the counter after being pushed. In Video B, the massive 100 KG iron block remains completely stationary [. . .] The binary difference is unambiguous."
Judge stricter Veo 3.1 Material/Medium	Dry vs. soaking-wet cotton t-shirt, lifted by the arm	1 / 0	"Neither video shows any sagging; both shirts remain completely rigid and flat while suspended. The expected binary difference in sag and fall behavior due to water weight is not observable."
Judge more lenient Seedance 2.0 Spatial Alignment	Gripper stops above the plate vs. halts early over the bare table	0 / 1	"In Video A, the donut is released and lands on the red plate, whereas in Video B, the donut is released and lands on the bare wooden table. The final states match the expected divergence."

the image editor used for the edited frames (Gemini 3 Pro Image) is not the judge model. This comparison does not establish the absence of vendor bias.

Examples. Table 21 shows four O_p checks from run 1, one for each combination of judge verdict and human label, taken from tuples where all annotators agreed (chosen by a seeded shuffle). On the 811 paired checks where the annotators agree with each other, the judge agrees with them in 87.7% of cases, and its disagreements lean toward being stricter (59 against 41).

Controls. As a control on false passes we judged 15 Veo 3.1 pairs (one per sub-category configuration, drawn at random) after replacing each pair by two samples of the same prompt (runs 1 and 2) or by the two videos swapped. O_p passes in 13% of these control pairs, against 33% for the same 15 pairs judged as they are. This small control does not establish a calibrated false-positive floor for the full benchmark.

pAvg combines whether the pair changes correctly (A_p, P_p, O_p) with whether the rest stays the same (E_p). Without E_p, closed-source models average 36.0% and open-source models 17.0% (with E_p: 43.8% and 28.6%), and the ranking barely changes (Spearman 0.97).

E.8 VLM Judge Example Observations

Table 22 shows representative judge observations for several failure modes identified in Appendix C.2. All excerpts come from run 1.

First, the judge's language is highly specific—it does not simply say "the video is bad" but names the mechanism of failure ("the robotic arm exhibits severe morphing," "the gap increases instead," "a donut spontaneously spawns mid-video"). This specificity lets the judge distinguish between an action that was never executed (A failure), physics that was individually plausible but identical across the pair (P_p mode collapse), and an environment that hallucinated divergent artifacts (E_p divergence).

Second, the failure descriptions are qualitatively different between single and paired modes. Single-video failures describe generation-quality problems (morphing, teleportation, spawning), the kinds of artifacts that single-video benchmarks already target. Paired failures describe *relational* problems—the two videos are individually fine but fail to differ in the right way. This distinction supports the core premise of our benchmark: paired evaluation tests a property that single-video physics and environment checks cannot, namely contrastive reasoning.

Table 22: **Representative VLM judge observations by failure mode.** Each row shows one verbatim excerpt illustrating the failure pattern. Observations are lightly truncated for space.

Dim.	Mode	Judge Observation (verbatim excerpt)
A_s	Action Omission	*In Video A, the ego car maintains a constant speed and the gap to the front car remains stable throughout the video. In Video B, the ego car fails to accelerate; its speed remains constant and the gap to the front car does not decrease at any point, constituting a clear Action Omission.*
A_p	Baseline Inconsist.	*There is no baseline consistency between the two videos. The arm kinematics and trajectories are completely different from the start, and neither video successfully demonstrates the intended destination accuracy task. The spatial difference is not clearly or unambiguously [. . .]*
P_s	Object Morphing	*In Video A, the robotic arm exhibits severe morphing and kinematic violations, completely losing its structural integrity and blocking the camera view. In Video B, a donut spontaneously spawns mid-video, and a duplicate donut appears during the grasp, violating physical realism [. . .]*
P_p	Mode Collapse	*The expected trajectory bifurcation for a destination accuracy error should occur during the delivery phase after pick-up. However, the trajectories in both videos remain identical throughout the entire movement, with no divergence at the delivery phase.*
E_s	Background Collapse	*In Video A, the background and environment remain stable. In Video B, there is a severe background collapse at the end of the video (00:03-00:04), where the buildings on the left melt and stretch into horizontal lines, and the road surface loses its texture.*
E_p	Envir. Divergence	*Due to the severe camera shift and the spontaneous appearance of the wall logo in Video B, the workspace background and camera angle are completely inconsistent between the two videos.*
O_p	Mode Collapse	*The expected outcome is that Video A successfully grasps and lifts the dark cap, while Video B misses. However, in both videos, the dark cap remains on the bed. Video A fails to grasp and lift the cap, resulting in mode collapse where both videos show the same final object state*
O_p	Reversed outcome	*At no point does the gap in Video B become smaller than in Video A. A causal inversion occurs where the gap in Video B is much larger than in Video A by the end of the clip, meaning the intended outcome failed completely.*
O_p	Upstream Propag.	*The expected outcome is that Video A shows the donut placed on the red plate, while Video B shows it placed on the bare table. Neither video shows the donut being placed at any destination, resulting in a complete failure to produce the expected state divergence.*

F Experimental Setup and Statistical Reliability

Pass rates pool valid evaluations across the three generation runs unless stated otherwise. The human-agreement study and the GPT-5 judge use run 1 only.

F.1 Compute Resources and Implementation Details

Hardware. All open-source video generation was performed on a single workstation running Ubuntu 22.04 with 4× NVIDIA RTX A6000 GPUs (48 GB VRAM each). Each generation job was scheduled to a single GPU; we did not use tensor- or pipeline-parallel inference within a model.

Inference settings. For every open-source model we used the publicly released default inference configuration—sampler, number of denoising steps, classifier-free guidance scale, and the model's default precision. No model was fine-tuned, distilled, or adapted to the benchmark, and no per-prompt hyperparameter tuning was performed.

API-accessed components. The five closed-source generation models (Grok Imagine, Veo 3.1, Seedance 2.0, Seedance 1.5, Kling 3.0) were accessed through hosted APIs via kie.ai (Kie.ai,

Table 23: **Run-to-run variation.** pAvg (%) in each run, the mean (the leaderboard value, up to rounding) and the standard deviation across runs (pp); right: SD across runs of the four paired components (pp). * closed-source.

	pAvg per run					SD across runs (pp)			
Model	Run 1	Run 2	Run 3	Mean	SD	A_p	P_p	E_p	O_p
CogVideoX1.5-5B	22.7	23.3	25.2	23.7	1.27	0.9	2.5	1.2	0.8
Wan2.2-5B	28.4	26.8	31.4	28.9	2.34	0.7	1.1	8.0	1.4
HunyuanVideo1.5-8.3B	27.0	29.4	26.5	27.6	1.54	2.0	1.4	2.5	1.3
Cosmos-Predict2-2B	32.9	33.9	35.7	34.2	1.39	1.3	1.2	2.7	0.8
Seedance 1.5*	32.4	29.7	27.8	30.0	2.33	3.0	1.5	3.7	1.9
Kling 3.0*	38.6	40.8	41.6	40.4	1.54	1.0	2.2	2.2	1.7
Seedance 2.0*	41.8	45.5	45.2	44.2	2.04	1.6	2.3	5.5	1.6
Veo 3.1*	50.9	51.7	51.6	51.4	0.48	2.2	2.4	3.8	1.7
Grok Imagine*	51.7	52.0	55.1	52.9	1.92	1.3	1.1	4.0	2.8

2026), and each pair $(\mathcal{V}_+, \mathcal{V}_-)$ was scored by Gemini 3.1 Pro as a VLM judge over the four APEO dimensions, also via API. We therefore consumed no local GPU resources for either component, and end-to-end latency was governed by vendor-side queueing.

Repeated generation runs. Runs 2 and 3 of the closed-source models were generated with the same endpoints and settings as run 1 (480p where the API exposes it; clips are 4 or 5 s long, depending on what each model supports). We set no seed for the closed-source models, so a re-run is a resubmission of identical inputs that draws fresh noise on the vendor side. The open-source models were re-run with a different fixed seed each time (CogVideoX1.5: 500 and 900; Cosmos-Predict2: 300 and 700; HunyuanVideo1.5: 1234 and 2025, with seed 42 in run 1; Wan2.2: 100 and 999). For every model we compared every video across runs by SHA-256 and found no identical pair.

F.2 Repeated Generation Runs

Protocol. Every model generated the benchmark three times. Grok scores use a fixed 317-pair subset across runs, excluding two pairs initially blocked by safety filtering (one each in Road Friction and Road Medium); the affected sub-category rates therefore use 14 pairs per run. Kling uses 319/316/316 pairs across runs after safety filtering (21 Grasp Align and 18 Push/Yank pairs in runs 2–3); other models use 319 per run. Aggregate rates pool valid pair evaluations across runs, weighting run-level rates by their valid sample counts.

The closed-source models are run without a seed, so a re-run repeats the request; the open-source models use a different fixed seed per run (Appendix F.1).

Spread across runs. Table 23 gives pAvg in each run and the standard deviation across runs (sample SD, n=3). The SD of pAvg is between 0.48 and 2.34 pp. E_p is the noisiest component (Wan2.2 SD 8.0, Seedance 2.0 SD 5.5).

Rankings. The top model is the same in all three runs overall and on five of the six primitives; on Surface Friction, Seedance 2.0 is first in run 2 (Table 24). Kendall's τ between the full orderings of two runs is 0.93 overall and 0.78–0.85 on four primitives, but 0.58 on Surface Friction and 0.66 on Obstacle Configuration.

Signal versus noise. For two models, the signal is the difference of their three-run mean pAvg and the noise is $\sqrt{\sigma_i^2 + \sigma_j^2}$ with σ the across-run SD. For the 20 closed-versus-open pairs (Table 25) the ratio is 0.3 to 20.4, median 7.2. The 3 pairs below 2 all involve Seedance 1.5, whose pAvg lies inside the open-source range. Grok and Veo differ by 1.5 pp against noise of 2.0 pp and form a joint top tier. Among the three leaders, Seedance 2.0 is 7 to 9 pp behind Grok and Veo on pAvg, several times its

Table 24: **Ranking stability across runs.** Best model by pAvg in each run, and the mean Kendall τ between the nine-model orderings of pairs of runs.

Scope	Run 1	Run 2	Run 3	mean τ
Overall pAvg	Grok	Grok	Grok	0.93
Spatial Alignment	Veo 3.1	Veo 3.1	Veo 3.1	0.85
Force/Degree	Grok	Grok	Grok	0.85
Temporal Sequencing	Veo 3.1	Veo 3.1	Veo 3.1	0.80
Surface Friction	Grok	Seedance 2.0	Grok	0.58
Material/Medium	Veo 3.1	Veo 3.1	Veo 3.1	0.78
Obstacle Configuration	Grok	Grok	Grok	0.66

Table 25: **Signal-to-noise ratio for closed-versus-open pairs**: absolute difference of three-run mean pAvg divided by $\sqrt{\sigma_i^2 + \sigma_j^2}$.

	CogVideoX	Wan2.2	HunyuanVideo	Cosmos
Seedance 1.5*	2.4	0.3	0.8	1.5
Kling 3.0*	8.3	4.1	5.8	3.0
Seedance 2.0*	8.5	4.9	6.5	4.1
Veo 3.1*	20.4	9.4	14.7	11.8
Grok Imagine*	12.7	7.9	10.3	7.9

run-to-run SD (2.0 pp), although the three are within 0.9 pp on sAvg (SD 0.3 to 1.6 pp; Table 27); it also has the largest gap between sAvg and pAvg of all nine models (25.9 pp).

F.3 Confidence Intervals

Method. Pairs from one scene share a frame, and in the driving longitudinal and overtaking sub-categories several pairs share one baseline video. We therefore resample scene clusters: pairs that share a baseline video form one cluster and every other pair is its own (256 clusters from 319 pairs). A cluster is drawn with replacement as a whole, keeping all models' results together; statistics are means over pairs, and each pair's value is first averaged over the three runs. Point estimates follow the reported aggregate rates; intervals use this scene-level resampling procedure. We use 10,000 resamples and percentile 95% intervals. For a difference between two groups of pairs, each group is resampled separately, except Grok's Longitudinal Acceleration minus Braking, whose two groups share the same 18 scenes and are resampled jointly. The four headline intervals also exclude zero at the 99% level, stricter than a Bonferroni correction over four quantities.

Headline effects. Table 26 shows the four headline quantities. The contrastive bottleneck holds for every model: $P_s - P_p$ is between 28 and 49 pp, smallest for Veo 3.1 and Grok Imagine and largest for HunyuanVideo, with every interval far from zero (Table 27, which also gives each model's all-four pass rate and sAvg run-to-run standard deviation).

Per-primitive scores. Spatial Alignment and Force/Degree clearly separate from Temporal Sequencing, Surface Friction and Obstacle Configuration; the O_p intervals of the latter three overlap, so we treat them as one hard group (Table 28).

Rank stability under resampling. In each draw we re-rank the nine models and check whether the full-data leader is still first (Table 29). Grok stays first in 81% of draws on pAvg and Veo in 65% on O_p, as expected for a near tie; on the primitives the leader is retained in 84–99.9% of draws. The top-three set is stable except on Material/Medium and Obstacle Configuration (Jaccard overlap 0.78 and 0.72), where third place is contested.

Sub-categories. Sub-category estimates rest on 14–30 pairs and have correspondingly wider intervals (for example Object Mass O_p = 7.9% [3.5, 13.1]).

Table 26: **Headline effects with 95% and 99% cluster-bootstrap intervals** (percentage points).

Quantity	Estimate	95% CI	99% CI	Pairs
Closed-source minus open-source pAvg (pp)	15.2	[13.3, 17.1]	[12.7, 17.8]	319
Contrastive bottleneck, HunyuanVideo1.5: $P_s - P_p$ (pp)	49.3	[44.4, 53.8]	[43.0, 55.2]	319
Force/Degree minus Surface Friction, O_p (pp)	25.2	[17.7, 32.6]	[15.6, 35.2]	52 + 30
Grok Imagine: Longitudinal Accel. minus Braking, O_p (pp)	81.5	[68.5, 92.6]	[63.0, 96.3]	18 (same scenes)

Table 27: **Per-model statistics quoted in Section 4** (three generation runs). All four: share of pairs that pass A_p, P_p, E_p and O_p. sAvg SD: standard deviation of sAvg across the three runs. $P_s - P_p$: difference of the Table 2 entries, with its 95% scene-cluster bootstrap interval. * closed-source.

Model	All four (%)	sAvg SD (pp)	$P_s - P_p$ (pp) [95% CI]
CogVideoX1.5	6.2	2.6	33.3 [27.6, 38.7]
Wan2.2	8.5	3.1	39.2 [34.5, 43.7]
HunyuanVideo1.5	5.4	3.3	49.3 [44.4, 53.8]
Cosmos-Predict2	11.1	2.2	37.7 [33.2, 42.0]
Seedance 1.5*	10.4	2.8	37.5 [33.3, 41.2]
Kling 3.0*	16.1	2.0	41.9 [37.5, 45.8]
Seedance 2.0*	21.7	1.6	42.3 [38.1, 46.3]
Veo 3.1*	26.9	0.3	27.7 [23.6, 31.5]
Grok Imagine*	28.9	0.6	29.5 [25.6, 33.2]

Table 28: **Per-primitive scores (%) and 95% cluster-bootstrap intervals**, mean over models.

Primitive	Pairs	pAvg [95% CI]	O_p [95% CI]
Spatial Alignment	95	42.9 [40.5, 45.3]	32.2 [29.2, 35.3]
Force/Degree	52	44.7 [38.9, 50.4]	37.9 [31.4, 44.4]
Temporal Sequencing	82	31.2 [29.4, 33.1]	15.9 [13.5, 19.0]
Surface Friction	30	27.4 [24.3, 30.5]	12.7 [9.3, 16.5]
Material/Medium	31	41.0 [35.6, 46.2]	25.3 [20.1, 30.8]
Obstacle Configuration	29	26.5 [22.7, 30.5]	14.0 [10.5, 17.6]

Table 29: **Ranking stability under scene resampling.** Top-1 retention: share of draws in which the full-data leader is still first; τ: mean Kendall correlation with the full-data ordering.

Scope	Leader	Top-1 retention (%)	mean τ
Overall pAvg	Grok	81.4	0.96
Overall O_p	Veo 3.1	65.2	0.92
Spatial Alignment	Veo 3.1	90.3	0.92
Force/Degree	Grok	93.9	0.92
Temporal Sequencing	Veo 3.1	85.1	0.91
Surface Friction	Grok	84.2	0.86
Material/Medium	Veo 3.1	98.5	0.89
Obstacle Configuration	Grok	99.9	0.86

F.4 Release and Licenses

The repository holds the 319 prompt pairs and reference images, the videos of all nine models from all three runs, the item-level judge outputs with the judge's explanations, the SHA-256 report, the failure-mode labels, the audit files, the judge scripts, and the aggregation scripts that rebuild every table, so all numbers can be recomputed without generating a video. Our own material (prompts, judge outputs, scripts, metadata) is released under CC BY 4.0; the reference frames come from nuScenes (CC BY-NC-SA 4.0) and DROID (Apache 2.0) and keep those terms. Wan2.2-5B and CogVideoX1.5-5B are under Apache 2.0, Cosmos-Predict2-2B under the NVIDIA Open Model License, and HunyuanVideo1.5-8.3B under the Tencent Hunyuan Community License. The closed-source models were used through their public APIs (through kie.ai); their outputs are subject to each vendor's terms and are redistributed for academic evaluation only.

G Per-Primitive and Per-Domain Results

G.1 Per-Primitive Summary

We decompose the aggregate scores by the six physical primitives (See Appendix Tables 30–35); values below are means over the nine models.

On O_p, Force/Degree (37.9%) and Spatial Alignment (32.2%) are the easiest, Material/Medium follows (25.3%), and Temporal Sequencing (15.9%), Obstacle Configuration (14.0%) and Surface Friction (12.7%) overlap within their confidence intervals (Appendix F.3). Each primitive pools a driving and a robotic task that can differ a lot (Appendix G.3); Force/Degree, for example, is easy mainly because of acceleration in driving.

Temporal Sequencing is hard, mainly in driving. Even the best model (Veo 3.1) reaches only $O_p = 29.7\%$, and the closed-source advantage is smallest here (7.5 pp on pAvg vs. 22.2 pp on Force/Degree). A representative judge observation on an overtaking pair is that *"the relative position between vehicles is only marginally different between the two timing conditions"*. The difficulty is concentrated in driving (O_p 12.3% against 25.8% in manipulation).

G.2 Per-primitive Analysis Results

Tables 30–35 break down each model's scores by physical primitive (pooled over the three generation runs). We take the three action primitives first, then the three environment ones.

Spatial Alignment (Table 30). This primitive tests whether models can generate different outcomes when the *position* of an action changes—for example, whether the ego car accelerates in the correct lane, or whether the gripper closes at the right spot. Veo 3.1 leads with 61.5% pAvg, closely followed by Grok Imagine at 57.9%. The P_s-to-P_p drop is moderate here compared to other primitives: the top two models retain about two-thirds to three-quarters of their single-video physics scores when evaluated as pairs, so spatial differences are among the easier ones for models to render. Open-source models lag substantially, with CogVideoX1.5 at just 30.2%.

Force/Degree (Table 31). This is the easiest primitive on average, but that is mostly driving (O_p 47.6% against 22.4% in manipulation; Appendix G.3). It asks whether models can show different *magnitudes* of the same action—gentle vs. hard braking, or a light push vs. a forceful yank. Grok Imagine reaches 68.1% pAvg here, the highest score any model gets on any primitive. A possible explanation for the high score is that differences in acceleration or force produce visually obvious motion changes (the car moves much faster or much slower). This is also where the closed-source vs. open-source gap is largest (22.2 pp): the top two closed-source models both exceed 62%, while no open-source model breaks 42%.

Temporal Sequencing (Table 32). This primitive is about *when* sub-actions happen relative to each other—for example, whether the ego car accelerates before or after a gap opens up, or whether the gripper closes before or after touching the object. It is one of the hardest primitives, and the range across models is narrow: pAvg runs from 18.2% (Seedance 1.5) to 42.9% (Veo 3.1), a 25 pp spread against 42 pp on Force/Degree. P_p is low for every model (at most 24.0%), so few trajectory

Table 30: **Spatial Alignment** — per-model results (%), pooled over three runs. Columns as in Table 2.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
CogVideoX1.5	37.7	24.9	41.9	24.6	25.4	52.6	18.6	30.2
Wan2.2	40.2	27.7	55.1	27.7	49.6	63.9	23.9	35.8
HunyuanVideo1.5	42.1	21.4	66.5	27.0	50.4	64.6	15.4	32.1
Cosmos-Predict2	41.1	29.1	65.3	35.1	60.9	76.8	21.1	40.5
Seedance 1.5*	51.6	36.5	56.7	25.6	38.6	41.1	29.5	33.2
Kling 3.0*	57.1	36.7	76.0	42.0	66.6	74.9	28.6	45.6
Seedance 2.0*	64.7	45.6	81.1	46.3	71.9	65.3	39.6	49.2
Veo 3.1*	71.4	57.5	78.4	59.3	73.0	69.8	59.3	61.5
Grok Imagine*	63.3	52.6	78.8	52.6	74.2	72.3	54.0	57.9

Table 31: **Force/Degree** — per-model results (%), pooled over three runs. Columns as in Table 2.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
CogVideoX1.5	39.4	17.9	57.4	14.1	31.4	57.1	16.0	26.3
Wan2.2	40.4	20.5	59.9	19.2	43.3	66.7	21.8	32.1
HunyuanVideo1.5	42.9	16.0	71.5	13.5	45.8	70.5	17.3	29.3
Cosmos-Predict2	51.0	27.6	70.8	27.6	57.4	84.0	28.2	41.8
Seedance 1.5*	60.3	41.0	61.9	35.3	40.4	53.8	48.7	44.7
Kling 3.0*	52.6	32.9	75.0	31.6	65.5	76.3	33.6	43.6
Seedance 2.0*	66.3	48.1	81.4	42.9	73.4	77.6	44.9	53.4
Veo 3.1*	68.9	63.5	72.8	55.1	67.3	70.5	62.2	62.8
Grok Imagine*	77.2	70.5	75.3	60.9	67.3	72.4	68.6	68.1

pairs diverge at the right moment. This is consistent with the idea that coordinating the timing of one agent's action relative to another agent's state is a different challenge from simply generating different action intensities.

Surface Friction (Table 33). Here, the action is the same in both prompts (e.g., hard braking), and only the surface property changes (dry asphalt vs. icy road, or smooth glass vs. rubber mat). This turns out to be very hard: the best O_p is just 37.9% (Grok Imagine), and eight of the nine models are below 20%. P_p is especially low: Cosmos-Predict2 scores $P_p = 2.2\%$, so it almost never produces a physically correct difference in how friction affects the outcome. A likely difficulty is that friction has no direct visual signature: even when the frame shows snow, the model has to know how snow changes braking, and a dry and an icy road can look alike until the car actually brakes.

Table 32: **Temporal Sequencing** — per-model results (%), pooled over three runs. Columns as in Table 2.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
CogVideoX1.5	37.2	10.6	65.4	7.3	33.7	69.1	11.0	24.5
Wan2.2	43.1	12.2	68.9	9.3	53.5	72.0	12.6	26.5
HunyuanVideo1.5	36.8	11.4	73.2	8.5	57.1	78.5	7.3	26.4
Cosmos-Predict2	32.1	14.6	62.4	10.2	63.0	87.8	10.2	30.7
Seedance 1.5*	30.9	7.7	61.8	6.5	42.7	50.0	8.5	18.2
Kling 3.0*	52.6	21.1	72.0	14.6	68.9	83.7	16.7	34.0
Seedance 2.0*	56.5	29.3	79.5	22.8	75.6	73.2	27.2	38.1
Veo 3.1*	64.0	34.1	76.0	24.0	78.7	83.7	29.7	42.9
Grok Imagine*	57.7	27.2	66.1	19.9	71.1	90.2	19.9	39.3

Table 33: **Surface Friction** — per-model results (%), pooled over three runs. Columns as in Table 2.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
CogVideoX1.5	26.1	22.2	15.0	3.3	14.4	23.3	5.6	13.6
Wan2.2	24.4	25.6	32.8	3.3	45.0	45.6	4.4	19.7
HunyuanVideo1.5	42.8	35.6	46.1	4.4	39.4	36.7	5.6	20.6
Cosmos-Predict2	42.2	28.9	30.0	2.2	35.6	44.4	2.2	19.4
Seedance 1.5*	58.9	41.1	50.0	8.9	39.4	34.4	11.1	23.9
Kling 3.0*	63.9	46.7	51.7	14.4	66.1	70.0	17.8	37.2
Seedance 2.0*	68.3	52.2	57.2	14.4	73.9	75.6	18.9	40.3
Veo 3.1*	55.0	42.2	35.0	12.2	39.4	38.9	11.1	26.1
Grok Imagine*	70.7	55.2	65.5	33.3	61.5	55.2	37.9	45.4

Table 34: **Material/Medium** — per-model results (%), pooled over three runs. Columns as in Table 2.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
CogVideoX1.5	30.1	26.9	26.9	2.2	37.6	45.2	3.2	19.4
Wan2.2	43.0	35.5	38.7	8.6	44.6	66.7	10.8	30.4
HunyuanVideo1.5	46.8	34.4	45.7	8.6	55.4	73.1	11.8	32.0
Cosmos-Predict2	64.0	40.9	46.2	16.1	55.9	73.1	16.1	36.6
Seedance 1.5*	62.4	41.9	55.9	22.6	59.7	76.3	30.1	42.7
Kling 3.0*	60.2	54.8	57.5	22.6	75.3	82.8	20.4	45.2
Seedance 2.0*	57.0	50.5	54.3	18.3	81.2	84.9	18.3	43.0
Veo 3.1*	78.0	66.7	73.7	63.4	74.2	62.4	71.0	65.9
Grok Imagine*	71.1	51.1	68.9	41.1	70.0	76.7	45.6	53.6

Material/Medium (Table 34). This primitive changes the material properties of the objects or the medium the agent operates in—for example, asphalt vs. a muddy road for driving, or a rigid wooden block vs. a soft sponge for manipulation. Veo 3.1 stands out here with 65.9% pAvg and $O_p = 71.0\%$, well ahead of Grok Imagine (53.6%). One possible explanation is that material differences are at least partially visible (wood looks different from sponge), giving models a visual cue to work with—unlike friction, whose effect on the motion is much harder to see. A_p is also at least 50% for the top four models, so they render distinct actions on the two materials even when P_p is lower.

Obstacle Configuration (Table 35). This primitive tests whether models react differently when the scene layout changes—a clear road vs. one with a traffic cone, or a light object vs. one too heavy to push. Grok Imagine dominates at 54.6% pAvg, about 21 pp ahead of the next model (Seedance 2.0 at 33.9%). E_p is low: across models, it averages 42.4% on this primitive, the lowest of the six (Surface Friction is next at about 47%, the others between 65% and 77%), likely because changing the obstacle configuration changes what the scene looks like, making it harder for models to keep the non-obstacle parts of the background identical across the pair. HunyuanVideo1.5 and CogVideoX1.5 both score $P_p = O_p = 1.1\%$, so they almost never respond correctly to the obstacle change. The primitive mixes a driving task (Scene Layout, O_p 21.2%) and a robotic one (Object Mass: 14 scenes, Appendix D.5; 6.3%), and its low average comes mostly from Object Mass.

G.3 Primitive Scores by Domain

Each primitive pools a driving task and a robotic one, and they are different tasks. Temporal Sequencing is overtaking timing in driving and the order of gripper actions in manipulation. Obstacle Configuration is a cone or a pothole in the lane in driving and an object too heavy to push in manipulation. Table 36 splits the primitive scores by domain.

Three primitives differ a lot between the two domains. Force/Degree is much easier in driving (O_p 47.6%) than in manipulation (22.4%). Temporal Sequencing goes the other way: 12.3% in driving against 25.8% in manipulation. Obstacle Configuration is 21.2% in driving (Scene Layout) and only 6.3% in manipulation (Object Mass, 14 scenes). Material/Medium also differs (20.4% in driving,

Table 35: **Obstacle Configuration** — per-model results (%), pooled over three runs. Columns as in Table 2.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
CogVideoX1.5	39.1	16.1	29.9	1.1	13.8	25.3	1.1	10.9
Wan2.2	40.2	20.7	43.1	4.6	46.6	32.2	3.4	15.2
HunyuanVideo1.5	47.1	26.4	44.3	1.1	40.2	35.6	1.1	16.1
Cosmos-Predict2	37.9	20.7	39.1	9.2	60.9	56.3	2.3	22.1
Seedance 1.5*	54.6	28.7	38.5	10.3	38.5	26.4	11.5	19.3
Kling 3.0*	62.6	46.0	49.4	14.9	62.1	62.1	11.5	33.6
Seedance 2.0*	65.5	47.1	58.0	18.4	68.4	55.2	14.9	33.9
Veo 3.1*	66.1	40.2	54.6	26.4	63.2	35.6	28.7	32.8
Grok Imagine*	79.9	65.5	67.8	48.3	67.2	52.9	51.7	54.6

29.9% in manipulation), while Spatial Alignment and Surface Friction are similar in the two domains. The hardest cells are therefore driving Temporal Sequencing, robotic Object Mass, and Surface Friction in both domains.

Table 36: **Primitive scores by domain** (mean over the nine models and three runs, %). AD = Autonomous Driving, Rob = Robotic Manipulation. Pairs = number of pairs in each domain (AD / Rob), counted as in Table 15.

		pAvg		O_p	
Primitive	Pairs	AD	Rob	AD	Rob
Spatial Alignment	53 / 42	42.9	42.9	32.5	31.9
Force/Degree	32 / 20	52.3	32.3	47.6	22.4
Temporal Sequencing	60 / 22	30.3	33.7	12.3	25.8
Surface Friction	15 / 15	29.4	25.4	13.7	11.9
Material/Medium	15 / 16	43.4	38.7	20.4	29.9
Obstacle Configuration	15 / 14	28.6	24.2	21.2	6.3

G.4 AD vs. Robotics Domain Comparison

Table 37 splits each model's paired scores by domain: Autonomous Driving (AD) and Robotic Manipulation (Rob).

Seven of the nine models score higher on AD than on Robotics. The gap is largest for CogVideoX1.5 (+14.3 pp) and Seedance 1.5 (+12.2 pp), suggesting these models' apparent capabilities are heavily biased toward road scenarios. However, Cosmos-Predict2 shows the *opposite* pattern: it scores 14.5 pp higher on Robotics than AD.

Veo 3.1 is almost perfectly domain-balanced (Δ = +0.4 pp), and Grok Imagine is close (+4.3 pp); Kling 3.0 (+1.0 pp) and HunyuanVideo1.5 (-0.8 pp) are balanced too, but at lower scores. The two top models achieve similar performance across the two evaluated domains.

E_p (environment consistency) is higher in AD than in Robotics for eight of the nine models (Veo 3.1 is the exception), likely because driving scenes have simpler, more predictable backgrounds (roads, lane markings) compared to cluttered tabletop manipulation environments.

G.5 Detailed Per-Primitive Results

Tables 38–44 give the full APEO breakdown per model and sub-category.

Each table is organized by sub-category (italic headers, with the number of pairs). Bold values indicate the best score per column within each sub-category.

The Object Mass block lists all 15 pairs; Table 35 uses 14 (Appendix D.5).

Table 37: **Per-model pAvg scores (%) split by domain** (pooled over three generation runs). AD = Autonomous Driving (11 sub-categories); Rob = Robotic Manipulation (7 sub-categories). Δ = AD $-$ Rob. * = closed-source.

	AD				Robotics				
Model	A_p	P_p	E_p	pAvg	A_p	P_p	E_p	pAvg	ΔpAvg
CogVideoX1.5	24.0	14.2	64.6	29.5	12.1	9.0	32.6	15.2	+14.3
Wan2.2	24.9	15.6	67.5	31.5	18.9	15.0	54.0	25.1	+6.4
HunyuanVideo1.5	20.0	10.0	70.5	27.3	22.5	19.4	56.1	28.1	-0.8
Cosmos-Predict2	17.2	9.6	76.7	28.3	37.7	35.7	73.9	42.8	-14.5
Seedance 1.5*	37.7	21.2	54.7	34.9	18.9	15.8	35.4	22.7	+12.2
Kling 3.0*	36.8	20.5	85.6	40.7	33.9	34.9	63.0	39.8	+1.0
Seedance 2.0*	47.7	31.4	75.4	46.6	36.2	31.5	65.1	40.6	+6.0
Veo 3.1*	52.1	40.9	63.5	51.6	47.8	45.0	71.6	51.2	+0.4
Grok Imagine*	51.6	41.5	77.0	54.7	48.3	43.4	69.8	50.4	+4.3

- **Extreme task-level variance within a single model.** Grok Imagine scores $A_p = 75.9\%$ on Longitudinal Acceleration but $A_p = 11.1\%$ on Longitudinal Braking—within the same Spatial Alignment primitive. Similarly, Veo 3.1 achieves $O_p = 98.1\%$ on Longitudinal Acceleration but $O_p = 11.1\%$ on Overtake Accel.
- P_s **can be near-perfect while** P_p **is close to zero.** HunyuanVideo1.5 achieves $P_s = 93.9\%$ on Overtake Hard (almost every video is plausible on its own) but $P_p = 1.1\%$ (almost no pair diverges as physics predicts). This is the "individually plausible, but unresponsive to the intervention" failure mode at its most extreme.

Table 38: **Spatial Alignment (Autonomous Driving) — per-sub-category results (%)** (pooled over three generation runs). Full APEO breakdown for each model on every AD sub-category within this primitive. **Bold**: best per column within each sub-category.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
Long. Accel (AD), $n = 18$ pairs								
CogVideoX1.5	66.7	48.1	80.6	44.4	30.6	48.1	48.1	47.2
Wan2.2	73.1	59.3	79.6	50.0	65.7	72.2	68.5	62.5
HunyuanVideo1.5	53.7	20.4	94.4	20.4	56.5	68.5	20.4	32.4
Cosmos-Predict2	43.5	16.7	77.8	16.7	60.2	75.9	18.5	31.9
Seedance 1.5*	60.2	57.4	66.7	31.5	44.4	46.3	57.4	48.1
Kling 3.0*	63.9	37.0	96.3	33.3	72.2	72.2	37.0	44.9
Seedance 2.0*	74.1	51.9	**97.2**	46.3	71.3	61.1	51.9	52.8
Veo 3.1*	**93.5**	**92.6**	88.0	**90.7**	66.7	61.1	**98.1**	**85.6**
Grok Imagine*	76.9	75.9	81.5	74.1	**79.6**	**79.6**	92.6	80.6
Long. Brake (AD), $n = 18$ pairs								
CogVideoX1.5	49.1	22.2	77.8	20.4	34.3	55.6	27.8	31.5
Wan2.2	40.7	14.8	86.1	16.7	75.0	72.2	20.4	31.0
HunyuanVideo1.5	44.4	11.1	94.4	11.1	67.6	77.8	11.1	27.8
Cosmos-Predict2	13.9	5.6	73.1	3.7	63.9	75.9	5.6	22.7
Seedance 1.5*	61.1	40.7	79.6	29.6	48.1	51.9	42.6	41.2
Kling 3.0*	50.9	14.8	95.4	14.8	76.9	**90.7**	16.7	34.3
Seedance 2.0*	63.9	35.2	**99.1**	31.5	79.6	72.2	42.6	45.4
Veo 3.1*	**72.2**	**46.3**	81.5	**46.3**	74.1	53.7	**46.3**	**48.1**
Grok Imagine*	28.7	11.1	86.1	5.6	**84.3**	63.0	11.1	22.7
Lateral Avoid (AD), $n = 17$ pairs								
CogVideoX1.5	15.7	17.6	10.8	17.6	20.6	78.4	0.0	28.4
Wan2.2	18.6	17.6	10.8	15.7	29.4	66.7	0.0	25.0
HunyuanVideo1.5	29.4	25.5	14.7	19.6	28.4	60.8	0.0	26.5
Cosmos-Predict2	26.5	27.5	19.6	19.6	54.9	**86.3**	2.0	33.8
Seedance 1.5*	34.3	37.3	25.5	27.5	17.6	47.1	7.8	29.9
Kling 3.0*	60.8	51.0	43.1	51.0	**65.7**	84.3	15.7	50.5
Seedance 2.0*	60.8	64.7	45.1	**64.7**	53.9	84.3	27.5	60.3
Veo 3.1*	52.9	31.4	47.1	31.4	50.0	80.4	43.1	46.6
Grok Imagine*	**63.7**	**68.6**	**53.9**	**64.7**	50.0	70.6	**56.9**	**65.2**

Table 39: **Spatial Alignment (Robotic Manipulation) — per-sub-category results (%)** (pooled over three generation runs). Full APEO breakdown for each model on every Rob sub-category within this primitive. **Bold**: best per column within each sub-category.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
Release Coords (Rob), $n = 20$ pairs								
CogVideoX1.5	20.0	16.7	28.3	28.3	27.5	71.7	11.7	32.1
Wan2.2	33.3	26.7	49.2	30.0	37.5	71.7	16.7	36.2
HunyuanVideo1.5	40.0	25.0	67.5	46.7	55.0	80.0	21.7	43.3
Cosmos-Predict2	53.3	48.3	80.0	68.3	70.8	**96.7**	31.7	61.2
Seedance 1.5*	58.3	38.3	61.7	26.7	48.3	50.0	23.3	34.6
Kling 3.0*	46.7	36.7	76.7	58.3	64.2	80.0	33.3	52.1
Seedance 2.0*	71.7	48.3	85.8	53.3	77.5	70.0	50.0	55.4
Veo 3.1*	72.5	70.0	**88.3**	71.7	**87.5**	81.7	61.7	71.3
Grok Imagine*	**82.5**	**71.7**	86.7	**76.7**	80.8	83.3	**68.3**	**75.0**
Grasp Align (Rob), $n = 22$ pairs								
CogVideoX1.5	37.9	21.2	17.4	13.6	15.9	16.7	7.6	14.8
Wan2.2	35.6	21.2	49.2	25.8	42.4	40.9	15.2	25.8
HunyuanVideo1.5	42.4	24.2	59.8	33.3	43.9	39.4	21.2	29.5
Cosmos-Predict2	61.4	42.4	70.5	**57.6**	54.5	53.0	40.9	48.5
Seedance 1.5*	43.9	13.6	49.2	15.2	33.3	15.2	18.2	15.5
Kling 3.0*	63.3	43.8	68.0	50.0	56.2	51.6	37.5	45.7
Seedance 2.0*	54.5	31.8	76.5	37.9	75.0	43.9	27.3	35.2
Veo 3.1*	**65.9**	**47.0**	**83.3**	54.5	**81.8**	**71.2**	**48.5**	**55.3**
Grok Imagine*	62.9	37.9	82.6	42.4	74.2	65.2	42.4	47.0

Table 40: **Force/Degree — per-sub-category results (%)** (pooled over three generation runs). Full APEO breakdown for each model on every sub-category within this primitive. **Bold**: best per column within each sub-category.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
Brake Intensity (AD), $n = 18$ pairs								
CogVideoX1.5	45.4	9.3	88.0	7.4	49.1	55.6	9.3	20.4
Wan2.2	34.3	0.0	72.2	0.0	57.4	50.0	1.9	13.0
HunyuanVideo1.5	49.1	3.7	**93.5**	3.7	59.3	**74.1**	5.6	21.8
Cosmos-Predict2	51.9	9.3	88.0	7.4	74.1	72.2	13.0	25.5
Seedance 1.5*	54.6	35.2	71.3	33.3	46.3	51.9	37.0	39.4
Kling 3.0*	53.7	18.5	81.5	14.8	76.9	70.4	18.5	30.6
Seedance 2.0*	57.4	25.9	88.0	20.4	76.9	68.5	25.9	35.2
Veo 3.1*	75.0	59.3	88.0	55.6	**82.4**	57.4	63.0	58.8
Grok Imagine*	**83.3**	**74.1**	87.0	**68.5**	77.8	55.6	**81.5**	**69.9**
Accel Intensity (AD), $n = 14$ pairs								
CogVideoX1.5	53.6	31.0	91.7	35.7	46.4	97.6	35.7	50.0
Wan2.2	66.7	57.1	86.9	59.5	56.0	97.6	64.3	69.6
HunyuanVideo1.5	52.4	33.3	86.9	31.0	53.6	88.1	38.1	47.6
Cosmos-Predict2	72.6	59.5	77.4	57.1	52.4	95.2	64.3	69.0
Seedance 1.5*	79.8	76.2	75.0	57.1	46.4	88.1	78.6	75.0
Kling 3.0*	72.6	64.3	78.6	59.5	66.7	**100.0**	73.8	74.4
Seedance 2.0*	**100.0**	**100.0**	**97.6**	**97.6**	**84.5**	97.6	**100.0**	**98.8**
Veo 3.1*	84.5	97.6	84.5	85.7	57.1	90.5	97.6	92.9
Grok Imagine*	94.0	97.6	89.3	85.7	65.5	**100.0**	97.6	95.2
Push/Yank (Rob), $n = 20$ pairs								
CogVideoX1.5	24.2	16.7	5.8	5.0	5.0	30.0	8.3	15.0
Wan2.2	27.5	13.3	30.0	8.3	21.7	60.0	10.0	22.9
HunyuanVideo1.5	30.8	15.0	40.8	10.0	28.3	55.0	13.3	23.3
Cosmos-Predict2	35.0	21.7	50.8	25.0	45.8	**86.7**	16.7	37.5
Seedance 1.5*	51.7	21.7	44.2	21.7	30.8	31.7	**38.3**	28.3
Kling 3.0*	36.6	23.2	**66.1**	26.8	53.6	64.3	17.9	33.0
Seedance 2.0*	50.8	31.7	64.2	25.0	**62.5**	71.7	23.3	37.9
Veo 3.1*	52.5	43.3	50.8	33.3	60.8	68.3	36.7	45.4
Grok Imagine*	**60.0**	**48.3**	55.0	**36.7**	59.2	68.3	36.7	**47.5**

Table 41: **Temporal Sequencing — per-sub-category results (%)** (pooled over three generation runs). Full APEO breakdown for each model on every sub-category within this primitive. **Bold**: best per column within each sub-category.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
Overtake Hard (AD), $n = 30$ pairs								
CogVideoX1.5	41.1	1.1	90.0	1.1	42.8	94.4	1.1	24.4
Wan2.2	39.4	6.7	87.2	0.0	66.7	94.4	5.6	26.7
HunyuanVideo1.5	32.2	1.1	**93.9**	1.1	67.2	95.6	1.1	24.7
Cosmos-Predict2	18.3	1.1	80.0	1.1	65.6	**96.7**	1.1	25.0
Seedance 1.5*	24.4	7.8	72.2	3.3	46.1	71.1	6.7	22.2
Kling 3.0*	43.9	5.6	81.7	3.3	70.6	91.1	4.4	26.1
Seedance 2.0*	51.1	17.8	**93.9**	11.1	**75.6**	74.4	14.4	29.4
Veo 3.1*	**68.3**	**43.3**	85.6	**23.3**	75.0	87.8	**40.0**	**48.6**
Grok Imagine*	60.0	32.2	79.4	16.7	74.4	94.4	27.8	42.8
Overtake Accel (AD), $n = 15$ pairs								
CogVideoX1.5	44.4	20.0	**82.2**	20.0	44.4	86.7	22.2	37.2
Wan2.2	40.0	11.1	62.2	15.6	60.0	86.7	11.1	31.1
HunyuanVideo1.5	32.2	8.9	71.1	13.3	62.2	88.9	11.1	30.6
Cosmos-Predict2	10.0	0.0	50.0	2.2	54.4	**97.8**	2.2	25.6
Seedance 1.5*	20.0	8.9	56.7	8.9	46.7	60.0	8.9	21.7
Kling 3.0*	45.6	13.3	56.7	13.3	70.0	95.6	13.3	33.9
Seedance 2.0*	**52.2**	**28.9**	65.6	**28.9**	77.8	77.8	**31.1**	**41.7**
Veo 3.1*	51.1	11.1	55.6	8.9	**78.9**	91.1	11.1	30.6
Grok Imagine*	35.6	0.0	42.2	0.0	52.2	91.1	0.0	22.8
Overtake Brake (AD), $n = 15$ pairs								
CogVideoX1.5	40.0	17.8	74.4	8.9	40.0	84.4	17.8	32.2
Wan2.2	36.7	11.1	**77.8**	13.3	67.8	84.4	13.3	30.6
HunyuanVideo1.5	21.1	4.4	66.7	6.7	65.6	86.7	4.4	25.6
Cosmos-Predict2	17.8	4.4	47.8	4.4	77.8	91.1	6.7	26.7
Seedance 1.5*	21.1	6.7	67.8	13.3	60.0	57.8	8.9	21.7
Kling 3.0*	43.3	17.8	62.2	8.9	70.0	91.1	13.3	32.8
Seedance 2.0*	43.3	22.2	71.1	**20.0**	77.8	73.3	**26.7**	35.6
Veo 3.1*	**58.9**	**26.7**	62.2	15.6	**84.4**	80.0	24.4	**36.7**
Grok Imagine*	46.7	11.1	48.9	11.1	75.6	**100.0**	11.1	33.3
Gripper Timing (Rob), $n = 22$ pairs								
CogVideoX1.5	25.0	12.1	14.4	6.1	9.8	12.1	12.1	10.6
Wan2.2	54.5	21.2	42.4	15.2	21.2	22.7	22.7	20.5
HunyuanVideo1.5	56.8	31.8	50.8	16.7	34.1	42.4	15.2	26.5
Cosmos-Predict2	75.8	**50.0**	56.8	31.8	55.3	66.7	30.3	44.7
Seedance 1.5*	53.8	7.6	47.0	4.5	23.5	9.1	10.6	8.0
Kling 3.0*	75.8	**50.0**	75.8	34.8	65.2	60.6	37.9	45.8
Seedance 2.0*	75.8	**50.0**	75.0	36.4	72.7	68.2	**42.4**	49.2
Veo 3.1*	70.5	42.4	**86.4**	40.9	**79.5**	75.8	31.8	47.7
Grok Imagine*	**77.3**	**50.0**	75.8	**43.9**	76.5	**77.3**	28.8	**50.0**

Table 42: **Surface Friction — per-sub-category results (%)** (pooled over three generation runs). **Bold**: best per column within each sub-category.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
Road Friction (AD), $n = 15$ pairs								
CogVideoX1.5	45.6	42.2	23.3	6.7	22.2	8.9	8.9	16.7
Wan2.2	21.1	35.6	14.4	4.4	46.7	8.9	6.7	13.9
HunyuanVideo1.5	60.0	55.6	31.1	4.4	42.2	11.1	6.7	19.4
Cosmos-Predict2	44.4	31.1	22.2	2.2	32.2	15.6	2.2	12.8
Seedance 1.5*	72.2	73.3	38.9	11.1	37.8	22.2	11.1	29.4
Kling 3.0*	82.2	84.4	48.9	15.6	92.2	84.4	17.8	50.6
Seedance 2.0*	**87.8**	**86.7**	51.1	13.3	**93.3**	**86.7**	17.8	**51.1**
Veo 3.1*	67.8	60.0	34.4	6.7	41.1	8.9	8.9	21.1
Grok Imagine*	84.5	76.2	**65.5**	**40.5**	67.9	38.1	**42.9**	49.4
Table Friction (Rob), $n = 15$ pairs								
CogVideoX1.5	6.7	2.2	6.7	0.0	6.7	37.8	2.2	10.6
Wan2.2	27.8	15.6	51.1	2.2	43.3	**82.2**	2.2	25.6
HunyuanVideo1.5	25.6	15.6	61.1	4.4	36.7	62.2	4.4	21.7
Cosmos-Predict2	40.0	26.7	37.8	2.2	38.9	73.3	2.2	26.1
Seedance 1.5*	45.6	8.9	61.1	6.7	41.1	46.7	11.1	18.3
Kling 3.0*	45.6	8.9	54.4	13.3	40.0	55.6	17.8	23.9
Seedance 2.0*	48.9	17.8	63.3	15.6	54.4	64.4	20.0	29.4
Veo 3.1*	42.2	24.4	35.6	17.8	37.8	68.9	13.3	31.1
Grok Imagine*	**57.8**	**35.6**	**65.6**	**26.7**	**55.6**	71.1	**33.3**	**41.7**

Table 43: **Material/Medium — per-sub-category results (%)** (pooled over three generation runs). **Bold**: best per column within each sub-category.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
Road Medium (AD), $n = 15$ pairs								
CogVideoX1.5	55.6	53.3	51.1	2.2	68.9	62.2	4.4	30.6
Wan2.2	55.6	64.4	52.2	8.9	67.8	71.1	11.1	38.9
HunyuanVideo1.5	55.6	55.6	51.1	6.7	62.2	77.8	6.7	36.7
Cosmos-Predict2	58.9	51.1	44.4	2.2	56.7	60.0	2.2	28.9
Seedance 1.5*	63.3	57.8	58.9	17.8	63.3	80.0	22.2	44.4
Kling 3.0*	57.8	**75.6**	55.6	11.1	**88.9**	**88.9**	11.1	46.7
Seedance 2.0*	53.3	66.7	51.1	2.2	**88.9**	**88.9**	2.2	40.0
Veo 3.1*	**82.2**	71.1	**76.7**	**66.7**	75.6	51.1	**75.6**	**66.1**
Grok Imagine*	72.6	59.5	67.9	38.1	82.1	88.1	47.6	58.3
Deformability (Rob), $n = 15$ pairs								
CogVideoX1.5	4.4	0.0	4.4	0.0	8.9	28.9	0.0	7.2
Wan2.2	28.9	8.9	24.4	6.7	22.2	60.0	11.1	21.7
HunyuanVideo1.5	36.7	15.6	41.1	11.1	46.7	66.7	17.8	27.8
Cosmos-Predict2	66.7	31.1	48.9	28.9	57.8	**84.4**	28.9	43.3
Seedance 1.5*	60.0	24.4	54.4	24.4	57.8	71.1	37.8	39.4
Kling 3.0*	60.0	31.1	57.8	31.1	63.3	75.6	31.1	42.2
Seedance 2.0*	57.8	31.1	55.6	31.1	**73.3**	80.0	31.1	43.3
Veo 3.1*	**72.2**	**62.2**	68.9	**57.8**	**73.3**	71.1	**64.4**	**63.9**
Grok Imagine*	68.9	42.2	**70.0**	42.2	58.9	64.4	46.7	48.9

Table 44: **Obstacle Configuration — per-sub-category results (%)** (pooled over three generation runs). **Bold**: best per column within each sub-category.

Model	A_s	A_p	P_s	P_p	E_s	E_p	O_p	pAvg
Scene Layout (AD), $n = 15$ pairs								
CogVideoX1.5	60.0	24.4	46.7	0.0	20.0	15.6	2.2	10.6
Wan2.2	50.0	17.8	44.4	2.2	52.2	15.6	4.4	10.0
HunyuanVideo1.5	62.2	24.4	50.0	0.0	40.0	22.2	0.0	11.7
Cosmos-Predict2	38.9	4.4	36.7	0.0	76.7	57.8	0.0	15.6
Seedance 1.5*	71.1	42.2	55.6	13.3	45.6	15.6	17.8	22.2
Kling 3.0*	81.1	62.2	56.7	15.6	**84.4**	**73.3**	15.6	41.7
Seedance 2.0*	80.0	62.2	64.4	28.9	72.2	51.1	28.9	42.8
Veo 3.1*	68.9	40.0	60.0	26.7	63.3	15.6	42.2	31.1
Grok Imagine*	**91.1**	**82.2**	**83.3**	**71.1**	73.3	55.6	**80.0**	**72.2**
Object Mass (Rob), $n = 15$ pairs								
CogVideoX1.5	17.8	8.9	11.1	4.4	6.7	35.6	2.2	12.8
Wan2.2	32.2	22.2	42.2	8.9	40.0	53.3	2.2	21.7
HunyuanVideo1.5	33.3	26.7	37.8	2.2	43.3	53.3	2.2	21.1
Cosmos-Predict2	41.1	37.8	41.1	20.0	42.2	57.8	6.7	30.6
Seedance 1.5*	40.0	17.8	21.1	11.1	31.1	42.2	6.7	19.4
Kling 3.0*	46.7	33.3	44.4	17.8	38.9	53.3	6.7	27.8
Seedance 2.0*	53.3	35.6	**53.3**	11.1	**65.6**	**62.2**	4.4	28.3
Veo 3.1*	65.6	42.2	52.2	**31.1**	63.3	60.0	**20.0**	**38.3**
Grok Imagine*	**68.9**	**48.9**	52.2	26.7	61.1	53.3	**20.0**	37.2