FROM SIGNAL FIDELITY TO FUTURE-STATE ALIGNMENT: A HIERARCHICAL EVALUATION OF VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation models are increasingly viewed as candidate world models, yet visually convincing rollouts do not necessarily predict how a depicted scene actually evolves. Existing evaluations often conflate frame-level signal similarity or no-reference video quality with predictive state understanding. This work introduces FPAbench (Fidelity, Plausibility, and Alignment benchmark), a hierarchical evaluation of video world models that separates three questions: Signal Fidelity (L1), measuring frame and motion-signal proximity to an observed future; Video Plausibility (L2), measuring the visual coherence, temporal smoothness, motion, and apparent physical validity of an unconstrained rollout; and Future-State Alignment (L3), measuring whether the same rollout preserves entities, trajectories, spatial interactions, and causal state evolution relative to the observed future. FPAbench evaluates 15 video generation systems on 2,000 five-frame sequences spanning 13 physical-event categories. Across every evaluated system, L3 is lower than L2, and the same pattern holds in 190 of 195 model-category comparisons. The gap is largest for state-transition events involving gravity, collisions, fluids, and deformation, whereas relatively stable scenes are easier. Motion dynamism does not ensure future-state alignment. These results show that signal fidelity and plausible video appearance are insufficient evidence of world-model capability, motivating hierarchical evaluation from signals to event-specific future states in observed scenes, not in plausible appearance alone.

INTRODUCTION

World models are intended to support prediction by representing how an environment changes over time (Ha & Schmidhuber, 2018; Hafner et al., 2023; Bruce et al., 2024). Early latent-dynamics and planning systems made this predictive role explicit through learned rollouts and imagined control (Hafner et al., 2019; Schrittwieser et al., 2020). Video generation models are increasingly viewed as candidate world models because an initial image can be extended into a visually convincing rollout (Ho et al., 2022; Blattmann et al., 2023; Yang et al., 2024): subjects remain recognizable, cameras appear stable, and frames exhibit familiar texture and motion. Figure 1 illustrates that visual credibility alone is not predictive success. A model may produce a smooth, recognizable continuation while missing an object's landing, interaction, or visible state transition.

This distinction exposes a gap in how video world models are commonly evaluated. Full-reference metrics measure whether generated frames are close to the observed sequence in signal space (Wang et al., 2004; 2003; Zhang et al., 2018; Teed & Deng, 2020). No-reference video metrics and visual-language judges measure whether a rollout looks coherent, temporally smooth, dynamic, and free of conspicuous artifacts (Unterthiner et al., 2018; Huang et al., 2024; Sun et al., 2025; Liu et al., 2024). Both are useful, but neither alone asks whether the generated event follows the future that was actually observed in the depicted scene. Consequently, a system can be rated highly for plausible appearance while failing at the state evolution that a predictive world model should capture.

This work therefore introduces FPAbench, a hierarchical evaluation that separates three levels of evidence for video world models. *Signal Fidelity* (L1) evaluates frame and motion-signal proximity to the observed future. *Video Plausibility* (L2) evaluates an AI-only rollout without reference to the

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

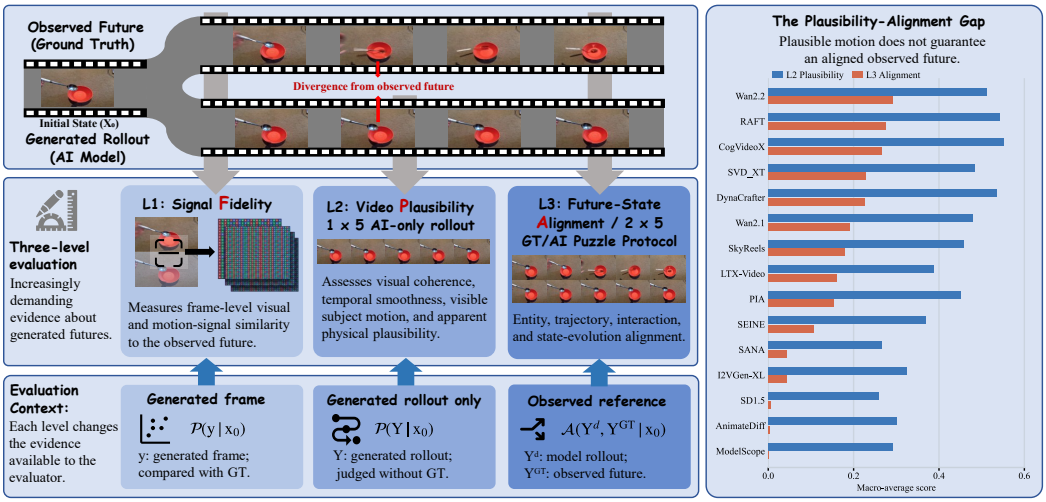


Figure 1: Future-state alignment exposes a gap hidden by visual plausibility. Left: a visually coherent rollout diverges from the observed final state. The three cards separate evidence about signal fidelity, no-reference video plausibility, and observed-future state alignment. Right: paired L2 and L3 averages for the 15 evaluated systems; every system has lower L3 than L2.

ground truth: does it look coherent, temporally smooth, dynamic, and apparently physically plausible? *Future-State Alignment* (L3) evaluates the same rollout against the observed future: does it preserve relevant entities, trajectories, spatial interactions, and causal state evolution? The hierarchy prevents a favorable answer to a lower-level question from being interpreted as evidence for a higher-level capability. It also treats the observed sequence as one realized future, rather than claiming that every alternative continuation is physically invalid in general.

FPAbench instantiates this evaluation on 2,000 five-frame sequences spanning 13 physical-event categories and 15 systems, from geometric extrapolation and early diffusion baselines to modern image-to-video generators. The evidence supports the need for the hierarchy at both model and event levels: all 15 evaluated systems receive lower Future-State Alignment scores than Video Plausibility, and 190 of 195 model-category pairs follow the same direction. The discrepancy is not simply a failure to generate motion. Motion Dynamism is high on average but negatively associated with model-level Future-State Alignment. The gap is especially pronounced for state-transition events, including gravity, collisions, fluids, and deformation, whereas stable visual scenes are easier.

FPAbench makes three contributions. First, it formulates a hierarchical evaluation that distinguishes Signal Fidelity, no-reference Video Plausibility, and observed-future Future-State Alignment as complementary evidence. Second, it provides a 2,000-sequence, 13-category diagnostic set and a structured Puzzle Protocol that separately evaluates the three levels. Third, its 15-system study shows that signal fidelity, visible motion, and plausible appearance are insufficient evidence that a generator predicts the observed state evolution of a physical event under evaluation.

Related Work

Video generation and image-to-video research have substantially improved visual fidelity, motion quality, and temporal coherence through diffusion and transformer-based systems (Ho et al., 2022; Singer et al., 2022; Blattmann et al., 2023; Zhang et al., 2023; Xing et al., 2024; Yang et al., 2024). Their evaluations commonly measure frame similarity, perceptual quality, distributional realism, generated motion, or physical commonsense (Unterthiner et al., 2018; Huang et al., 2024; Bansal et al., 2024; Meng et al., 2024). These signals are necessary for a useful generator, but they are often interpreted together as evidence that a rollout is successful. The premise of FPAbench is that these signals answer different questions: signal closeness and visual plausibility do not, by themselves, establish that a generated rollout predicts an observed physical event.

Benchmark	Scale	Observed future?	Physical-event taxonomy	Judge
No-reference evaluation of video generation (the judge never sees what happened next)				
VBench (Huang et al., 2024)	944 prompts, 16 dims	✗	—	metrics + human pref.
VideoPhy (Bansal et al., 2024)	688 captions	✗	material types	human + VideoCon-Physics
PhyGenBench (Meng et al., 2024)	160 prompts, 27 laws	✗	27 physical laws	VLM+LLM (PhyGenEval)
PhyWorldBench (Gu et al., 2026)	1,050 prompts	✗	motion / energy / rigid	zero-shot MLLM + human
WorldModelBench (Li et al., 2025)	350 inst., 7 domains	✗	physics adherence	2B judge + 67K human labels
Prediction compared against a ground-truth future (synthetic or simulated environments only)				
WorldBench (Upadhyay et al., 2026)	425 Kubric seqs. (9-frame input)	✓(sim.)	disentangled concepts	SAM2 masks vs. sim. + QA
Physion (Bear et al., 2021)	≈600 scen., 8 families	✓(synth.)	mechanical events	VQA + readout models
Physical reasoning / QA without generation (ground truth is a label)				
IntPhys (Riochet et al., 2018)	4 concept blocks	✗	block-world concepts	violation-of-expectation
CLEVRER (Yi et al., 2020)	10k synthetic videos	✓(synth.)	collision events	MCQ / descriptive
PHYRE (Bakhtin et al., 2019)	36 task templates	✗	2D mechanics	action correctness
PhysBench (Chow et al., 2025)	10k multimodal entries	✗	19 subclasses / 8 dims	75-VLM benchmark
FPAbench (ours)	2k real sequences	✓(real)	13 event cats., 3 groups	72B VLM, paired L2/L3

Table 1: Comparison with existing evaluation benchmarks. No-reference benchmarks judge arbitrarily long generated clips without a prescribed prediction horizon, whereas FPAbench requires every system to commit to the same five-frame horizon (one input frame and four observed future frames). To our knowledge, FPAbench is the only benchmark that evaluates open-domain image-to-video generation against an *observed real future*: no-reference benchmarks judge a rollout without access to what actually happened next, and benchmarks that do compare predictions against a ground-truth continuation operate only in synthetic or simulated environments.

Research on world models, physical prediction, and physical reasoning assesses state dynamics through interactive environments, synthetic physics, visual reasoning, and video-generation benchmarks (Bakhtin et al., 2019; Yi et al., 2020; Bear et al., 2021; Qin et al., 2024; Li et al., 2025; Gu et al., 2026; Upadhyay et al., 2026; Chow et al., 2025). These settings provide valuable tests of dynamics, but open-domain video generation from a single image is commonly assessed with a mixture of frame-level and appearance-level criteria. FPAbench complements them with an ordered diagnostic: it preserves lower-level measurements while requiring a separate observed-future comparison before a higher-level world-model claim is made. FPAbench neither proposes a new generator nor claims a short rollout defines a complete world model.

VLM-based evaluation has also become a practical way to assess multi-faceted visual and semantic properties (Bai et al., 2025). The contribution of FPAbench is not the use of a VLM judge alone, but its role in maintaining the separation of evaluation layers: the same rollout is assessed once without a reference future and once against its temporally aligned observed future. Throughout, FPAbench treats the single judge and lack of human calibration as limitations rather than evidence of validated reliability for these judgments. Table 1 surveys how existing benchmarks relate to FPAbench: no-reference protocols judge a generated video without access to what happens next, whereas benchmarks that do compare against a ground-truth future are confined to synthetic environments. In place of a closed ontology, the 13 FPAbench categories are constructed as diagnostic axes of event structure (stable, state-transition, and long-horizon causal events) designed to separate appearance preservation from visible event transitions, unlike quality- or coverage-oriented prompt suites.

HIERARCHICAL EVALUATION FRAMEWORK

Figure 2 summarizes the evaluation flow: the same standardized rollout is retained while the available evidence and scoring target become progressively more demanding from L1 through L3.

PROBLEM FORMULATION

Each evaluation instance begins with an initial frame x_0. A generator d produces a standardized five-frame rollout Y^d, which is evaluated against the temporally corresponding observed sequence Y^{GT}. The FPAbench framework does not collapse these observations into one score. Instead, it uses

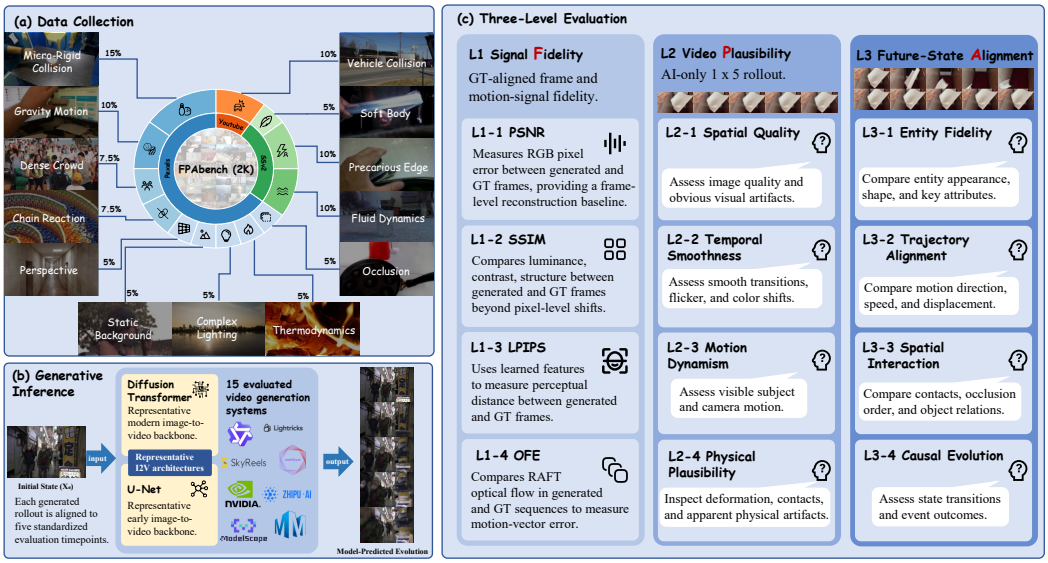


Figure 2: Hierarchical evaluation framework. The protocol uses 2,000 five-frame sequences across 13 physical-event categories and 15 systems under standardized rollout timepoints. It keeps Signal Fidelity, no-reference Video Plausibility, and observed-future Future-State Alignment as distinct layers through the Puzzle Protocol.

three ordered expressions to make explicit what evidence is available at each layer:

$$\mathcal{P}(\mathrm{y} \mid \mathrm{x}_0), \qquad \mathcal{P}(\mathrm{Y} \mid \mathrm{x}_0), \qquad \mathcal{A}(\mathrm{Y}^d, \mathrm{Y}^{\mathrm{GT}} \mid \mathrm{x}_0).$$

Here, y denotes a generated frame and Y denotes a generated rollout. The first expression concerns frame-level signal evidence; the second concerns an unconstrained rollout's appearance; the third evaluates the actual rollout of generator d against the observed future. $\mathcal{A}$ is an evaluation alignment score, not a likelihood assigned by the generator. Thus, a favorable L1 or L2 result remains informative but cannot by itself answer the L3 question.

The notation is a conceptual progression from signal association, through unconstrained rollout plausibility, to observed-future comparison. It does not instantiate a $do(\cdot)$ intervention or pose a strict counterfactual query. A single image can admit multiple physically plausible continuations; FPAbench therefore interprets Y^{GT} as one observed continuation and L3 as alignment to that continuation, not as a test of universal physical validity in this benchmark.

Dataset

The evaluation dataset contains 2,000 temporally ordered five-frame sequences. In each sequence, `frame_01` supplies x_0 and the subsequent frames constitute the observed future. The data are collected from public video sources, including Pexels, Something-Something V2 (Goyal et al., 2017), and YouTube. Rather than grouping clips only by visual topic, FPAbench organizes them by the physical-event structure that the future frames make observable. This taxonomy is part of the hierarchical design: it creates cases in which preserving short-term appearance and predicting a visible event transition are deliberately separated.

Each sequence retains an identifiable initial state and a temporally ordered future that exposes the relevant motion, interaction, or state change. FPAbench uses the first sampled frame as the only model input and preserves the four remaining frames as the reference rollout. The protocol therefore asks every model to commit to the same prediction horizon from the same visual evidence, while allowing each method to use its intended native inference configuration. It does not assume that the observed continuation is the unique physically valid one; it provides a concrete future against which event-specific alignment can be diagnosed.

The 13 categories form three analysis groups. Stable visual scenes contain Dense Crowd (150 sequences), Perspective (100), Static Background (100), and Complex Light (100). State-transition

events contain Fluid Dynamics (200), Gravity Motion (200), Micro Rigid (300), Occlusion (100), Soft Body (100), and Vehicle Collision (200). Long-horizon or causal events contain Chain Reaction (150), Precarious Edge (200), and Thermodynamics (100). The category labels are event-oriented rather than claims of complete physical coverage: for example, a falling or colliding object requires a trajectory and resulting state to align, whereas a stable scene can reward appearance preservation with limited state change. Table 3 reports all category counts and results.

Category sizes range from 100 to 300 sequences. For each model, FPAbench first averages rollouts within a category and then macro-averages the 13 category means. This prevents a large category such as Micro Rigid from dominating the overall result, while retaining the category-level values needed to expose where lower- and higher-level evidence diverge in practice across events.

The benchmark is intended as a diagnostic set for observed-future prediction, not as a claim to cover every real-world physical process. Every sequence was manually inspected before inclusion to verify that the initial state is identifiable, the event category label matches what the future frames make observable, and the frames meet the resolution and artifact requirements of the standardized rollout protocol. Sequences that failed these checks were removed during collection. FPAbench does not assume that publicly available source material is absent from a model's pretraining data, nor does FPAbench claim complete source-level deduplication without an auditable check. The supplementary Dataset Card will document source proportions, collection and selection procedures, category definitions, preprocessing, deduplication scope, known biases, and the license-dependent release plan for the benchmark.

THREE-LEVEL EVALUATION PROTOCOL

The layers are deliberately not score components to be averaged into a single "world-model score." They differ in reference information, evaluation target, and permitted claim. The hierarchy reports increasingly demanding evidence: frame or motion-signal proximity at L1, reference-free rollout plausibility at L2, and observed-future state alignment at L3.

L1: Signal Fidelity. L1 compares generated and ground-truth frames using PSNR, SSIM, LPIPS, and optical-flow endpoint error (OFE) (Wang et al., 2004; Zhang et al., 2018; Teed & Deng, 2020). These metrics measure frame-level visual or motion-signal proximity, averaged over the generated future frames. The four metrics deliberately cover complementary error modes: PSNR captures pixel-level reconstruction error, SSIM structural similarity, LPIPS learned perceptual distance, and OFE disagreement in motion signals. They are useful indicators of reconstruction fidelity, but they do not independently establish plausible dynamics or correct event evolution. For compact model comparison in Table 2, FPAbench reports an equal-weight L1 aggregate after aligning metric directions and normalizing each metric across the 15 evaluated systems. This aggregate is only a within-table signal-fidelity overview; it is not on the same absolute scale as L2 or L3 for cross-layer ranking.

L2: Video Plausibility. L2 presents a Qwen2.5-VL-72B judge with a 1×5 mosaic of the input frame followed by the generated frames at the four remaining evaluation timepoints, sampled uniformly over the source clip. Four independent calls score Spatial Quality, Temporal Smoothness, Motion Dynamism, and Physical Plausibility. The protocol asks whether a rollout appears visually coherent, temporally stable, dynamic, and free of obvious physical artifacts without exposing the reference future.

L3: Future-State Alignment. L3 presents the same judge with a 2×5 Puzzle Protocol mosaic: the observed ground-truth sequence and generated sequence are visually separated and temporally aligned, with the input frame as the first column of the ground-truth row. Four independent calls score Entity Fidelity, Trajectory Alignment, Spatial Interaction, and Causal Evolution. L3 therefore evaluates whether the rollout follows the observed future in motion, interaction, and state change. As defined in the Problem Formulation, a low L3 score is diagnostic evidence that the rollout does not reproduce the observed event, not a verdict that unobserved alternatives are physically wrong.

Both L2 and L3 use the same ordinal rubric: 1 for satisfactory outcomes, 0.5 for minor defects, and 0 for failure, followed by a macro-average over their four dimensions. The shared rubric is what makes the paired design interpretable: Δ is only meaningful when both layers are measured on the same scale, and L3's additional complexity is expressed through its dimensions and evidence rather than a harsher scale—its judge sees a reference-aligned 2×5 puzzle mosaic and scores

Model	Signal Fidelity (L1)					Video Plausibility (L2)					Future-State Alignment (L3)					Gap
Name	PSNR	SSIM	LPIPS	OFE	L1	Sp.	Temp.	Dyn.	Phys.	L2	Ent.	Traj.	Sp.	Caus.	L3	Δ
Geometric Extrapolation																
RAFT	15.29	0.545	0.393	136.5	0.934	0.357	**0.524**	0.457	**0.831**	0.542	0.351	0.265	0.296	0.188	0.275	0.267
Diffusion and Bridging Baselines																
SD1.5-img2img	14.60	0.478	0.544	156.3	0.724	0.023	0.026	**0.958**	0.027	0.258	0.012	0.002	0.009	0.004	0.007	0.252
AnimateDiff	11.19	0.431	0.714	264.5	0.272	0.256	0.031	0.753	0.161	0.300	0.005	0.002	0.004	0.002	0.003	0.297
PIA	14.72	0.530	0.481	168.0	0.797	0.294	0.357	0.640	0.515	0.452	0.219	0.126	0.179	0.094	0.154	0.297
Early Image-to-Video																
SVD_XT	14.59	0.535	0.448	155.0	0.835	0.283	0.417	0.642	0.589	0.483	0.288	0.208	0.255	0.160	0.227	0.255
I2VGen-XL	12.19	0.404	0.626	226.7	0.401	0.129	0.118	0.921	0.131	0.325	0.050	0.037	0.046	0.036	0.042	0.283
SEINE	14.10	0.518	0.488	160.0	0.765	0.188	0.221	0.802	0.268	0.370	0.145	0.099	0.110	0.073	0.107	0.263
DynamiCrafter	13.96	0.513	0.491	160.5	0.752	0.394	0.458	0.671	0.617	0.535	0.312	0.176	0.259	0.161	0.227	0.308
Modern Image-to-Video																
CogVideoX-5B	**15.72**	**0.580**	**0.383**	141.8	**0.983**	**0.402**	0.497	0.559	0.746	**0.551**	0.335	0.246	0.292	0.192	0.266	0.285
LTX-Video	14.60	0.522	0.463	164.1	0.800	0.201	0.290	0.609	0.451	0.388	0.204	0.144	0.181	0.113	0.161	0.227
Wan2.2-5B	15.42	0.557	0.389	**129.8**	0.962	0.342	0.439	0.564	0.700	0.511	**0.383**	**0.267**	**0.304**	**0.215**	**0.292**	**0.219**
Wan2.1-14B	13.53	0.485	0.512	170.6	0.681	0.278	0.350	0.850	0.438	0.479	0.255	0.156	0.196	0.152	0.190	0.289
SkyReels-1.3B	14.34	0.545	0.511	204.2	0.724	0.294	0.343	0.760	0.433	0.457	0.231	0.152	0.195	0.138	0.179	0.279
SANA-Video-2B	11.61	0.445	0.625	256.7	0.369	0.076	0.108	0.782	0.103	0.267	0.057	0.036	0.048	0.033	0.044	0.223
Text-to-Video Control																
ModelScopeT2V	10.17	0.298	0.796	305.8	0.000	0.247	0.011	0.756	0.153	0.292	0.001	0.000	0.001	0.001	0.001	0.291

Table 2: Layer-wise model results, averaged over each system's evaluated rollouts. L2 and L3 are means across their four listed dimensions. For L1, higher PSNR/SSIM and lower LPIPS/OFE are favorable; L1 is their direction-aligned, per-metric min–max normalized equal-weight mean, used only for within-table signal-fidelity comparison. $\Delta = \mathrm{L2} - \mathrm{L3}$ is the plausibility–alignment gap. Bold denotes the best value per column: LPIPS/OFE are minimized and Δ is closest to zero.

structured state-evolution dims, whereas L2 judges appearance-only dims on a reference-free rollout. The reference information and scoring targets remain intentionally different: L2 asks whether an unconstrained rollout looks acceptable, whereas L3 asks whether that rollout follows the observed event. They are therefore paired layers, not interchangeable sub-scores. FPAbench defines $\Delta = \mathrm{L2} - \mathrm{L3}$ as the operational plausibility-alignment gap: the positive shortfall exposed when the same rollout is assessed against its ground-truth-aligned observed future rather than only for no-reference plausibility. It is neither a causal-effect estimate nor an identical-task accuracy drop.

Puzzle Protocol and Judge Safeguards

The Puzzle Protocol operationalizes the L2–L3 separation without asking a judge to infer which frames belong together. L2 receives only the generated rollout. In L3, ground-truth frames appear above generated frames with a red separator and matched temporal positions. Each scoring dimension is evaluated in an independent API call, with chain-of-thought-guided structured JSON output and robust parsing. The evaluation pipeline records atomic JSONL outputs, retries failed requests, and uses resumable execution with adaptive request concurrency. These are safeguards for stable, auditable layer-specific execution; they do not constitute human validation of the judge. FPAbench uses a single Qwen2.5-VL-72B judge and does not perform a human calibration study in the current submission of this benchmark.

Experiment

Evaluation Setup

FPAbench evaluates 15 systems spanning a geometric extrapolation baseline, early diffusion and bridging baselines, early U-Net image-to-video systems, modern DiT image-to-video systems, and a text-to-video control. For image-conditioned systems, the framework supplies the same initial frame and aligns outputs to the same five-frame evaluation timepoints. Each system uses its official or author-recommended public inference configuration, fixed across the complete dataset; FPAbench does not tune settings by category, sample, L2 score, or L3 score. FPAbench does not force a single artificial native resolution across systems, because doing so can move a method outside its intended operating regime. The system set includes optical-flow extrapolation (Teed & Deng,

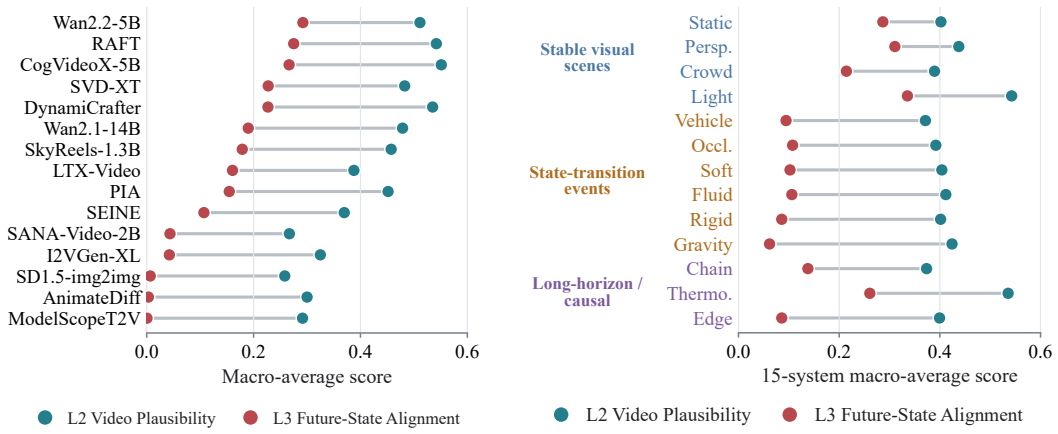


Figure 3: The plausibility–alignment gap. Left: model level. Each line connects a system's Video Plausibility (L2) and Future-State Alignment (L3) scores; models are ordered by L3. Every one of the 15 evaluated systems has L3 $<$ L2. Right: category level. Lines connect 15-system macro-average L2 and L3 scores; categories are ordered by $\Delta = \text{L2} - \text{L3}$.

2020), latent-diffusion baselines (Rombach et al., 2022; Guo et al., 2024), early image-to-video models (Blattmann et al., 2023; Zhang et al., 2023; Chen et al., 2023; Xing et al., 2024), and recent video generators (Yang et al., 2024; Wang et al., 2023). To preserve task comparability without imposing false hyperparameter equality, each image-conditioned system receives the same x_0 and is sampled at the target five evaluation positions. ModelScopeT2V remains a caption-conditioned control, while failed outputs are retained in reported coverage rather than selectively rerun.

Evaluated Systems. RAFT is a constant-flow extrapolation baseline that warps the input image with a fixed, linearly scaled motion field, estimated by RAFT on a synthetically shifted copy of that frame and using no observed future information (Teed & Deng, 2020). SD1.5-img2img is a latent-diffusion image-to-image baseline adapted to produce rollout frames (Rombach et al., 2022), and AnimateDiff augments a text-to-image diffusion backbone with learned motion modules (Guo et al., 2024). PIA uses plug-and-play modules for personalized image animation (Zhang et al., 2024). SVD_XT is an image-conditioned latent video diffusion model (Blattmann et al., 2023); I2VGen-XL uses cascaded diffusion for image-to-video synthesis (Zhang et al., 2023). SEINE targets video transition and prediction (Chen et al., 2023), while DynamiCrafter animates open-domain images with video diffusion priors (Xing et al., 2024). CogVideoX-5B is evaluated through its image-conditioned configuration (Yang et al., 2024). LTX-Video is a diffusion image-to-video model (Lightricks, 2024). Wan2.2-5B and Wan2.1-14B are Wan-family image-to-video variants (Team Wan et al., 2025). SkyReels-V2-I2V-1.3B targets long-form image-to-video generation (Chen et al., 2025a), and SANA-Video-2B is an efficient image-to-video diffusion transformer (Chen et al., 2025b). Finally, ModelScopeT2V is a text-to-video control that receives a caption derived from the input image rather than direct image conditioning (Wang et al., 2023). For each rollout, FPAbench averages the four dimensions within L2 or L3 and computes $\Delta_i = \text{L2}_i - \text{L3}_i$; it never averages L1, L2, and L3 into one score, and it

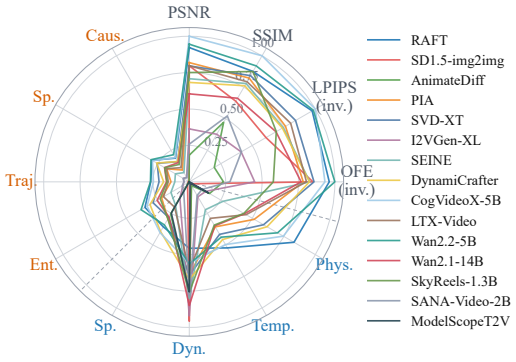


Figure 4: Layer-wise capability profiles across all 15 evaluated systems. Each curve is one system. L1 axes are direction-aligned, per-metric min–max normalized across systems; L2 and L3 axes retain their raw scores. The profiles are descriptive evidence by layer and do not define a cross-layer aggregate or overall ranking.

Event	N	L2	SD	Ent.	Traj.	Sp.	Caus.	L3	SD	Δ	Best sys.
Stable visual scenes											
Crowd	150	0.389	0.126	0.217	0.198	0.240	0.201	0.214	0.153	0.175	Wan2.2-5B
Persp.	100	0.438	0.137	0.304	**0.357**	0.296	0.286	0.311	0.217	0.127	RAFT
Static	100	0.402	0.111	0.279	0.296	0.289	0.283	0.287	0.213	**0.115**	Wan2.2-5B
Light	100	**0.543**	0.128	**0.329**	0.313	**0.347**	**0.353**	**0.336**	0.232	0.207	RAFT
State-transition events											
Fluid	200	0.412	0.098	0.164	0.087	0.124	0.048	0.106	0.081	0.306	DynamiCrafter
Gravity	200	0.424	0.102	0.122	0.032	0.080	0.013	0.062	0.042	0.362	DynamiCrafter
Rigid	300	0.402	0.103	0.166	0.054	0.096	0.028	0.086	0.063	0.316	DynamiCrafter
Occl.	100	0.392	0.102	0.155	0.093	0.122	0.059	0.107	0.083	0.285	CogVideoX-5B
Soft	100	0.404	0.100	0.163	0.081	0.120	0.046	0.102	0.080	0.302	Wan2.2-5B
Vehicle	200	0.371	0.108	0.136	0.085	0.117	0.039	0.094	0.084	0.277	Wan2.2-5B
Long-horizon / causal events											
Chain	150	0.374	0.118	0.192	0.122	0.152	0.085	0.138	0.113	0.236	Wan2.2-5B
Edge	200	0.399	0.108	0.163	0.059	0.099	0.023	0.086	0.066	0.313	Wan2.2-5B
Thermo.	100	0.536	0.154	0.276	0.231	0.260	0.276	0.261	0.219	0.275	Wan2.2-5B

Table 3: Category-level results, macro-averaged across 15 systems. Ent., Traj., Sp., and Caus. are the L3 Entity Fidelity, Trajectory Alignment, Spatial Interaction, and Causal Evolution components; L3 is their mean. SD is the between-system standard deviation of L2 or L3. Best sys. is the system with the highest L3 in that category. $\Delta = \mathrm{L2} - \mathrm{L3}$ and N is the category sequence count. Bold denotes the best value per column and the smallest gap.

macro-averages the 13 category means (Dataset section). The 15 systems are the fixed evaluated set, so $15/15$ and $190/195$ are coverage counts, not significance tests over a sampled model population. Aggregate confidence intervals use 10,000 category-stratified paired-bootstrap resamples of input sequence IDs; paired L2 and L3 are never resampled independently. The supplementary material records each system's version, entry point, preprocessing, native resolution, sampling settings, seed, valid coverage, and failure handling. Each input currently has one fixed-seed rollout, so the results characterize fixed configurations.

RQ1: Does Lower-Layer Plausibility Establish Future-State Alignment?

Figure 3 (left) and Table 2 show the separation: all systems have $\mathrm{L3} < \mathrm{L2}$, and 190 of 195 model-category pairs follow this direction. Treating a coherent rollout as event prediction would therefore overstate this fixed system set's observed-future capability.

This result does not imply that no model captures useful physical regularities. The stronger systems retain nonzero alignment scores and can preserve entities or portions of a short trajectory. Instead, the paired comparison shows why an unconstrained rollout is assessed more favorably when judged for visual plausibility than when it must match the observed continuation. Geometric extrapolation is a boundary case: short-term anchoring and stable backgrounds can support lower-layer signals without reliable event-specific prediction. The pattern is not driven by one model family: it persists for the optical-flow baseline, image-conditioned diffusion systems, and the text-to-video control. FPAbench therefore treats plausibility and observed-future alignment as paired, complementary diagnostics rather than interchangeable evidence of prediction.

RQ2: What Does Future-State Alignment Reveal Beyond Lower Layers?

Figure 4 provides the all-system layer-wise profile, while Table 3 exposes the upper-layer bottleneck: Entity Fidelity, Trajectory Alignment, Spatial Interaction, and Causal Evolution average .205, .154, .180, and .134, respectively. Causal Evolution is lowest. Visible motion is insufficient: L2 Motion Dynamism is .711 as the 13-category macro average (per-system values in Table 2 give .715), and it has a negative model-level association with L3 (Spearman $\rho = -.668$). Gravity Motion has Causal Evolution .013, while Fluid Dynamics, Micro Rigid, and Soft Body remain below .05. These errors are not confined to terminal frames: low trajectory alignment in the same categories shows that a recognizable rollout can commit to the wrong motion before the outcome. Causal Evolution captures

whether contact, transformation, or a resulting state matches the observed sequence, not merely whether motion is visible.

RQ3: Which Physical Events Make Layer Separation Most Necessary?

Figure 3 (right) and Table 3 localize the category-level evidence: stable visual scenes retain a mean L2 score of .443 and mean L3 score of .287, whereas state-transition events retain L2 .401 but fall to L3 .093, and the corresponding gap rises from .156 to .308. A favorable no-reference assessment can therefore conceal a failure to reproduce the observed trajectory, interaction, or resulting state.

Gravity Motion exemplifies the gap (.362): plausible motion can miss the expected contact, trajectory, or observed outcome. Table 4 identifies this as a trajectory–causal loss, motivating separate reporting of the layers rather than any combined single score (a claim about reporting, not that alternative continuations are physically impossible).

Table 3 localizes the same failure beyond the aggregate mean: Causal Evolution collapses in exactly the state-transition categories, with trajectory alignment similarly low.

Signal fidelity is likewise context, not alignment evidence: stronger frame-level structure and optical-flow signals retain aspects of the observed future but do not remove the paired L2–L3 gap, so FPAbench reports L1 only as a constrained descriptive overview in Table 2, reserving raw-metric correlations for supplementary analysis.

	Stable	State	Long hor.
L2	**0.443**	0.401	0.436
Traj.	**0.291**	0.072	0.137
Caus.	**0.281**	0.039	0.128
L3	**0.287**	0.093	0.162
Δ	**0.156**	0.308	0.275
CI	$[.150, .162]$	$[.305, .312]$	$[.270, .280]$

Table 4: Event-group profiles, macro-averaged over categories and 15 systems. Traj. and Caus. are the L3 Trajectory Alignment and Causal Evolution group means; CI is the category-stratified paired-bootstrap 95% interval for Δ. Bold denotes the best group per row.

The group-level intervals make the separation concrete: stable scenes have $\Delta = .156$ (95% paired-bootstrap CI $[.150, .162]$), whereas state-transition and long-horizon causal events increase to .308 and .275. Rollouts that appear comparably plausible across groups can therefore fail when an event requires contact, a state change, or sustained directional commitment.

Limitations and Future Works

The framework does not establish that current video generators possess no physical knowledge. Their rollouts exhibit useful visual priors, short-term continuity, and sometimes partial preservation of entities or trajectories. The evidence from FPAbench supports a narrower but actionable conclusion: signal fidelity and unconstrained plausibility are lower-layer evidence, not sufficient evidence of event-specific prediction, so world-model evaluations should report the layers separately.

Several limitations delimit this conclusion. First, L2 and L3 rely on a single Qwen2.5-VL-72B judge and have not been calibrated with a human study. Second, L3 compares against one observed future. Third, the 13 categories are imbalanced and may retain source or category biases despite macro-averaged reporting. Finally, systems differ in native resolution, inference configuration, and input adaptation; each input also uses one fixed-seed rollout. The evaluated set spans open-source and academic systems; commercial closed-source generators were not reachable under the standardized protocol and are left to a near-term extension with identical reporting.

Conclusion

This work introduces FPAbench, a hierarchical evaluation framework that separates Signal Fidelity, Video Plausibility, and Future-State Alignment. Across 15 systems and 13 physical-event categories, rollouts assessed as plausible without reference receive consistently lower scores when evaluated against the observed future. Future extensions should add multi-future coverage, human-calibrated assessment, and multi-seed analysis.

Reproducibility Statement

The complete evaluation pipeline, per-dimension score matrices for all 15 systems, judge prompt templates, and the benchmark metadata are included in the supplementary code-and-data archive. The supplementary technical document records each system's version, entry point, preprocessing, native resolution, sampling settings, seed, valid coverage, and failure handling, and documents the dataset sources, category definitions, and selection procedures. The per-category result matrices underlying every table in this paper can be regenerated from the released per-dimension outputs using the bundled aggregation scripts.

AI Use Statement

We used AI-assisted coding tools for routine engineering tasks during dataset preparation, evaluation-pipeline implementation, and the preparation of figures and tables; all scripts were executed and their outputs verified by the authors. We used large language models for language editing of the manuscript text. All scientific design choices, the hierarchical evaluation formulation, all experiments, and all reported conclusions were conducted and verified by the authors, who take full responsibility for the content of this work.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zhangping Fu, Yuanzhong Xu, Jiabo Ye, X.-C. Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

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

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiau-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin A. Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems*, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jake Bruce, Michael J. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy González, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zołna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025b.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, P. N. Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. Phyworldbench: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations*, 2026. Oral.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2555–2565, 2019.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.02060.

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

Lightricks. Ltx-video. Official model release, `https://huggingface.co/Lightricks/LTX-Video`, 2024.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond K. H. Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.02090.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

Yiran Qin et al. Worldsimbench: Towards video generation models as world simulators. In *Advances in Neural Information Processing Systems*, 2024.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588 (7839):604–609, 2020.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyang Hu, Harry Yang, Oron Ashual, Or Gafni, Devi Parikh, Sachin Gupta, Ishan Misra, and Yogesh Balaji. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pp. 402–419, 2020.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, pp. 1398–1402, 2003.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *Proceedings of the European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72952-2_23.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. 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 *International Conference on Learning Representations*, 2020.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7747–7756, 2024.