

4DWorldBench: A Comprehensive Evaluation Framework for 3D/4D World Generation Models

Yiting Lu^{1,*}, Wei Luo^{1,*}, Peiyan Tu^{2,3,*}, Haoran Li^{1,*}, Hanxin Zhu^{1,3,*}, Zihao Yu¹,
 Xingrui Wang¹, Xinyi Chen¹, Xinge Peng¹, Xin Li¹, Zhibo Chen^{1,3,†}

¹ University of Science and Technology of China

² Zhejiang University ³ Beijing Zhongguancun Academy

{luyt31415, lw21, lihr, hanxinzhu}@mail.ustc.edu.cn, pytu@zju.edu.cn
 {xin.li, chenzhibo}@ustc.edu.cn

Abstract

World Generation Models are emerging as a cornerstone of next-generation multimodal intelligence systems. Unlike traditional 2D visual generation, World Models aim to construct realistic, dynamic, and physically consistent 3D/4D worlds from images, videos, or text. These models not only need to produce high-fidelity visual content but also maintain coherence across space, time, physics, and instruction control, enabling applications in virtual reality, autonomous driving, embodied intelligence, and content creation. However, prior benchmarks emphasize different evaluation dimensions and lack a unified assessment of world-realism capability. To systematically evaluate World Models, we introduce the 4DWorldBench, which measures models across four key dimensions: Perceptual Quality, Condition-4D Alignment, Physical Realism, and 4D Consistency. The benchmark covers tasks such as Image-to-3D/4D, Video-to-4D, Text-to-3D/4D. Beyond these, we innovatively introduce adaptive conditioning across multiple modalities, which not only integrates but also extends traditional evaluation paradigms. To accommodate different modality-conditioned inputs, we map all modality conditions into a unified textual space during evaluation, and further integrate LLM-as-judge, MLLM-as-judge, and traditional network-based methods. This unified and adaptive design enables more comprehensive and consistent evaluation of alignment, physical realism, and cross-modal coherence. Preliminary human studies further demonstrate that our adaptive tool selection achieves closer agreement with subjective human judgments. We hope this benchmark will serve as a foundation for objective comparisons and improvements, accelerating the transition from "visual generation" to "world generation." Our project can be found at

<https://yeppp27.github.io/4DWorldBench.github.io/>.

1. Introduction

World Generation Models are rapidly emerging as a new paradigm for multimodal intelligence. In contrast to conventional visual generation, which typically optimizes frame fidelity or short-horizon temporal coherence, World Models aim to construct *realistic, dynamic, and physically consistent 3D/4D worlds* from text, image, or video conditions. Beyond high perceptual quality, World Models are expected to maintain *spatial-temporal coherence, physical realism, and interactive controllability*, enabling applications across *virtual reality and embodied AI*.

Existing benchmarks for generative models can be broadly categorized into three categories: video-centric, world generation-focused, and physics-aware. Video-based benchmarks (e.g., VBench2.0 [42]) emphasize high-level properties such as fidelity, creativity, commonsense, physics and controllability by leveraging broad prompt suites and hierarchical-level evaluations. However, they rely heavily on manually predefined question templates and offer only coarse evaluation of physical realism and 4D consistency. World generation benchmarks (e.g., WorldScore [10]) go beyond video generation to evaluate controllable 3D/4D generation across modalities, offering structured scene specifications and standardized metrics. While effective in assessing 3D/4D scene generation ability, they still exhibit limitations in fine-grained condition-4D alignment and provide insufficient coverage of physical realism. WorldSimBench [26] further extend evaluation to embodied world tasks by introducing a dual framework with explicit perceptual evaluation and implicit manipulative evaluation. Lastly, physics-aware benchmarks (e.g., PhyGenBench [24]) explicitly focus on testing adherence to physical laws through curated prompts and rule-based evaluation

* Equal contribution ‡Project leader † Corresponding author

pipelines. Meanwhile, Phyworld [17] investigates physical-law generalization in video generation models under controlled simulation settings. Although valuable for physical reasoning, these benchmarks often lack modality diversity, semantic richness, and scalability due to handcrafted logic. These gaps underscore the need for a more comprehensive, adaptive, and semantically grounded benchmark—such as our proposed 4DWorldBench—capable of evaluating 3D/4D world generation across diverse modalities and dimensions.

In short, the community lacks a *general-purpose, multimodal, and physics-aware* benchmark that measures *perceptual quality, condition-4D alignment, physical realism, and 4D consistency* in a single, extensible framework. In this work, we propose a more unified benchmark for world generation, named **4DWorldBench**. Our goal is to provide a comprehensive evaluation of *world generation models* that go beyond traditional visual synthesis, targeting the generation of coherent **3D and 4D worlds**. Specifically, 4DWorldBench supports models conditioned on multiple modalities, including **text, image, and video**, thereby enabling fair comparison across diverse input settings. Moreover, recognizing that world generation inherently involves not only perceptual fidelity but also compliance with the laws of nature, we explicitly introduce evaluation along the **physical realism dimension**, in addition to perceptual quality, condition-4D alignment, and 4D consistency. By combining these perspectives, 4DWorldBench aims to establish a more complete and challenging standard for assessing the next generation of world generation systems.

To support fine-grained evaluation across key dimensions—particularly condition-4D alignment and physical realism—we combine two complementary evaluation tracks. First, we employ an Multimodal Large Language Models (MLLM)-based Question Answer (QA) evaluator that inspects the generated video and answers a set of dimension-specific questions derived from the input condition (converted into text). Second, to handle more abstract reasoning—especially for assessing high-level physical realism—we introduce an LLM-based QA evaluator. This dual-track design is motivated by the observation that current MLLMs, while proficient at surface-level video question answering, often struggle with complex physical reasoning (e.g., fluid dynamics, force interactions). In contrast, LLMs excel at abstract reasoning and compositional understanding when provided with high-quality textual descriptions. By leveraging the strengths of both model types, our framework enables a more robust and semantically aware evaluation across a broad spectrum of world generation scenarios. Moreover, some dimensions need hybrid methods for evaluation: feature-based similarity metrics can be easily misled by videos containing only subtle object motions, resulting in inflated alignment scores; therefore, we addi-

Table 1. Comparison with representative benchmarks. “Modalities” refer to supported *condition* types. “Eval. dims” mark whether the benchmark covers Perceptual Quality (Q), Condition Alignment (A), Physical Realism (P), and 4D Consistency (C). “✓” indicates full support for the dimension, “partial” denotes limited or partially covered evaluation, and “–” indicates that the dimension is not supported.

Benchmark	Condition Modalities	Q	A	P	C
WorldScore [10]	Text	✓	partial	–	✓
VBench [16]	Text	✓	✓	partial	–
VBench2.0 [42]	Text / Image	✓	✓	partial	–
PhyGenBench [24]	Text	–	–	✓	–
4DWorldBench (ours)	Text / Image / Video	✓	✓	✓	✓

tionally incorporate an MLLM-based QA to guard against such deceptive cases. Our work offers three contributions:

- **Comprehensive, multimodal benchmark for world generation.** 4DWorldBench evaluates four main dimensions and multiple sub-dimensions, with balanced text/image/video condition sets for both non-physical and physical categories.
- **Adaptive hybrid evaluation methodology.** We integrate traditional metrics with LLM/MLLM-driven question decomposition and answering, enabling fine-grained, condition-aware assessment of alignment and physics while improving scalability.
- **Unified, extensible framework and leaderboard.** 4DWorldBench supports Image-to-3D/4D, Video-to-4D, Text-to-3D/4D, providing transparent, reproducible comparisons across models.

2. Related Work

2.1. Video-Based Evaluation Benchmarks

A large body of work focuses on evaluating video generation models in terms of fidelity, coherence, and semantic correctness. VBench [16] introduced a comprehensive benchmark that decomposes video quality into fine-grained dimensions such as identity consistency, motion smoothness, and temporal flicker, with scores aligned to human preferences. Its successor, VBench-2.0 [42], extends coverage to five key aspects: human fidelity, controllability, creativity, physics, and commonsense, leveraging vision-language models for deeper evaluation. Other works target specific challenges: DEVIL [21] evaluates visual vividness and the honesty of video content under dynamic settings; and T2V-CompBench [28] and TC-Bench [12] examine compositionality in multi-object, multi-attribute videos. Perceptual quality has also been tackled via large-scale frameworks such as EvalCrafter [23], which integrates objective metrics and human ratings into a unified leaderboard. Other benchmarks also highlight narrative or scene-level evaluation: StoryEval [32] measures story completion in multi-event prompts using VLM-based evaluators, showing that models struggle beyond a few coherent events. These efforts provide complementary views, but each fo-

cuses on a subset of evaluation dimensions.

2.2. 3D/4D World Generation Benchmarks

Evaluating 3D/4D generation requires capturing long-horizon spatial and temporal consistency. Benchmarks for text-to-3D scene generation: GT23D-Bench [5]) introduces a large-scale, well-organized 3D dataset with multimodal annotations and comprehensive 3D-aware metrics for evaluating text-3D alignment and visual quality, but are limited to static snapshots. And Eval3D [11] provides a fine-grained and interpretable evaluation framework that assesses 3D generation quality across semantic and geometric dimensions using consistency among diverse foundation models and tools. WorldScore [10] frames the task as sequential world generation with explicit instructions (e.g., camera trajectories), evaluating controllability, quality, and dynamics across multi-scene sequences. Overall, these benchmarks focus on measuring whether models can generate persistent, controllable, and temporally consistent 3D/4D worlds.

2.3. Physics-Focused Video Benchmarks

Recent advancements in video generation have led to the development of several benchmarks aimed at evaluating models’ understanding of physical dynamics. PhyGenBench [24] designs diverse prompts across multiple fundamental physical laws, including mechanics, optics, material science, and thermal properties. It provides a PhyGenEval pipeline that combines vision-language models and rule-based QA to assess physical plausibility. VideoPhy [3] and VideoPhy-2 [4] construct the dataset with explicit physical rules, testing models’ ability to understand physical interactions. PisaBench [18] focuses on real-world free-fall tasks to evaluate the ability of generative models to reproduce physical phenomena like gravity and dynamic collisions. LLMPhy [8] integrates large language models with physics engines for advanced physical reasoning, using datasets for predicting stable poses in complex object interactions. Physics-IQ [25] evaluates generative video models by giving them the initial frames of a real physical scene, having them predict the next few seconds, and then comparing the generated video against ground-truth recordings across multiple physics-based metrics.

3. Benchmark Dataset

To ensure comprehensive and controllable evaluation, we curate a benchmark dataset that spans both physics and non-physics domains, across text, image, and video modalities. The data collection process emphasizes category diversity, and physical plausibility.

Text Condition Collection. For non-physical modality, text prompts are sampled from WorldScore [10], a benchmark that decomposes world generation into next-scene prediction tasks and explicitly distinguishes between static and dynamic scenes. We sample the text condition across

several categories based on the structure and semantics of the text prompts: **Scene, Event, and Attribute**, which is shown in Fig. 7. For physical modality, to evaluate physical reasoning and control, we incorporate physics-driven prompts and exemplars following a structured taxonomy (see Fig. 7), which includes four major classes: Material, Mechanics, Optics, and Thermal phenomena. Each class is further divided into representative subclasses—for example, Optics includes refraction, reflection, and the Tyndall effect; Mechanics includes gravity, buoyancy, and pressure; Material covers color, hardness, combustibility, and redox reactions; and Thermal involves heating and phase transitions. These Text prompts are revised from PhyGenBench [24]. We further utilize GPT-5 to produce data augmentations for thermal physics. In total, the evaluation set includes 50 physical text conditions and 76 non-physical conditions in text form. **Image Condition Collection.** For non-physical condition, we sampled image conditions from Vbench2.0 [42]. We sample the image conditions from Vbench2.0 across several categories based on the structure and semantics of the captions: **Scene, Event, Object**. Specifically, captions involving interactive entities or motion are classified as **Event**, those describing a single interacting entity are labeled as **Object**, and captions without explicit entities are assigned to the **Scene** category. For object-level 3D generation model only, we sample 25 object-centric images (without background) from Objaverse-XL dataset [9]. For physical condition, We first collect videos from WISA [30], after which the first frame of each selected video is extracted to serve as the corresponding image conditioning input. The distribution of physical and non-physical image conditioning samples is shown in Fig. 3. Among them, 48 samples correspond to physical conditions, while 63 belong to non-physical scene conditions. **Video Condition Collection.** For non-physical condition, we source captioned video clips from VUDG [34]. These datasets provide diverse annotations, including scene type, weather condition, and crowd density. Our sampling strategy proceeds in two stages. First, we ensure coverage across stylistic and domain categories—such as cartoon, game, night, foggy, egocentric, shaky, and surveillance views. Second, within each category, we further stratify the samples by temporal motion amplitude, measured using inter-frame optical flow statistics. Accordingly, the videos are categorized into *low-motion* and *high-motion* types. This two-stage sampling process promotes domain diversity while ensuring comprehensive coverage of motion control capabilities.

For physical condition, we collect videos from WISA [30], which covers three physical domains—**Dynamic, Optic, and Thermodynamic**. We first uniformly sample videos from these three categories to ensure balanced domain coverage. Within each cat-

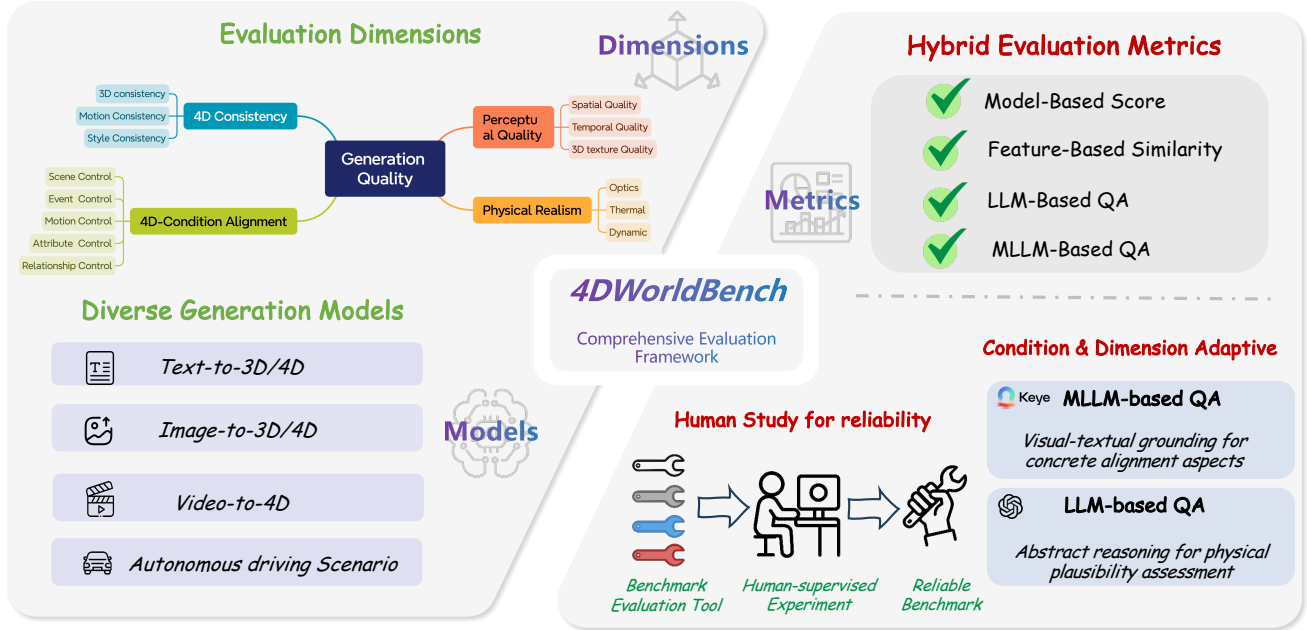


Figure 1. Overview of the 4DWorldBench framework. The benchmark evaluates generation quality across four key dimensions: 4D consistency, condition alignment, perceptual quality, and physical realism. It supports diverse generative settings, including text-, image-, and video-to-3D/4D generation. The framework integrates hybrid evaluation metrics—model-based scores, feature-based similarity, LLM-based QA, and MLLM-based QA—together with human studies for reliability. Condition- and dimension-adaptive QA modules leverage MLLMs for concrete visual grounding and LLMs for higher-level physical reasoning.

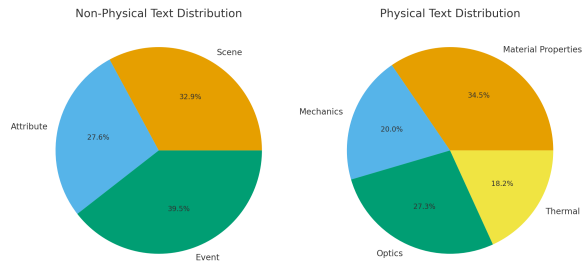


Figure 2. Text Condition Distribution.

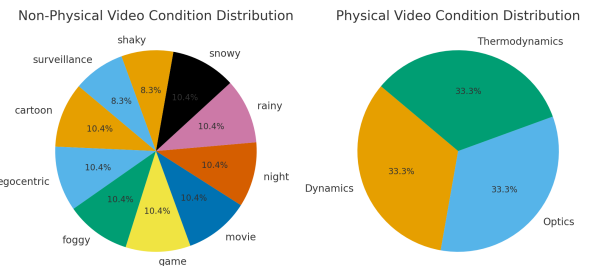


Figure 4. Video Condition Distribution.

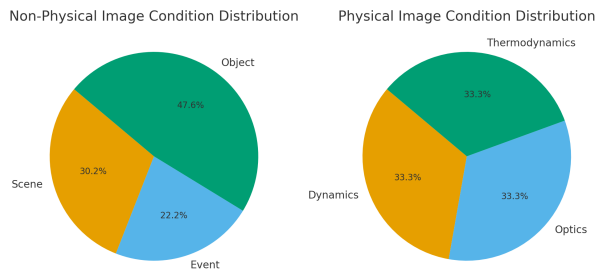


Figure 3. Image Condition Distribution.

egory, we further stratify the data into *high-motion* and *low-motion* subsets based on temporal motion amplitude. The distribution of physical and non-physical video condi-

tioning samples is shown in Fig. 4. Each subset contains 48 samples respectively.

4. Benchmark Metric

Our proposed 4DWorldBench includes four key dimensions: Physical Realism, Condition-4D Alignment, and 4D Consistency, Perceptual Quality.

4.1. Physical Realism

Our physical realism evaluation framework is built upon two core principles: (1) converting both the input conditions (except for text conditions) and the generated videos into captions, leveraging LLMs to assess physical consistency through caption-based reasoning instead of direct video observation; and (2) adaptively selecting evaluation dimensions based on the semantics of each input caption to ensure

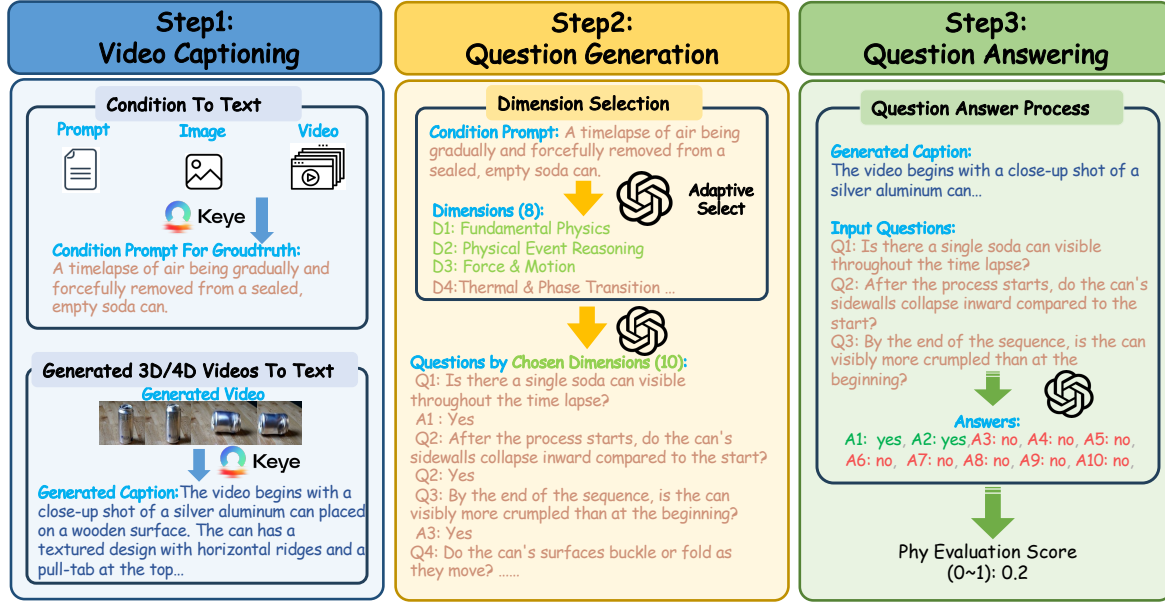


Figure 5. Pipeline of the Physical Realism Evaluation. The framework first unifies text, image, and video conditions into a common textual form, then uses an LLM to adaptively select relevant physical dimensions and generate diagnostic questions. An LLM-based QA module compares predicted and reference answers to yield a continuous physical realism score.

Table 2. Leaderboard for different categories of 4D generation models.

Model	Perceptual Quality			Condition-4D Alignment					Physical Realism			4D Consistency			Overall
	Spatial	Temporal	3D Texture	Event	Scene	Attribute	relationship	Motion	Dynamics	Optics	Thermal	Viewpoint	Motion	Style	
Image-to-4D															
Cam12V [43]	0.419±0.524	0.668	0.700	0.397	0.954	0.922	0.845	0.756	0.650	0.738	0.629	0.701	0.553±0.816	0.887	0.697
DiffusionAsShader [14]	0.509±0.522	0.790	0.715	0.400	0.965	0.963	0.804	0.781	0.688	0.850	0.763	0.908	0.874±0.750	0.918	0.763
Video-to-4D															
EX-4D [15]	0.305±0.409	0.356	0.616	0.615	0.805	0.792	0.879	0.869	0.421	0.627	0.600	0.266	0.384±0.811	0.827	0.599
ReCamMaster [2]	0.215±0.401	0.394	0.636	0.573	0.757	0.673	0.811	0.796	0.680	0.714	0.773	0.862	0.859±0.834	0.985	0.685
TrajectoryCrafter [39]	0.492±0.434	0.560	0.646	0.639	0.862	0.766	0.913	0.856	0.671	0.667	0.636	0.362	0.454±0.883	0.878	0.670
Vista [13]	0.216±0.402	0.397	0.657	0.405	0.686	0.516	0.698	0.607	0.607	0.613	0.593	0.921	0.942±0.622	0.986	0.617
Text-to-4D															
4Dfy [1]	0.336±0.469	0.442	0.569	0.566	0.505	0.326	0.553	0.545	0.265	0.417	0.200	0.741	0.934±0.705	0.993	0.535
dreamin4D [44]	0.234±0.396	0.181	0.655	0.409	0.576	0.390	0.558	0.416	0.235	0.238	0.150	0.612	0.761±0.360	0.850	0.439

scenario-relevant physical assessment. Specially, to evaluate the degree of physical realism in generated videos, we design a three-step framework that combines video captioning and condition captioning, physics-aware question generation, and question answering for quantitative scoring.

1) Video Captioning for Condition and Generated Video. The input condition C , originating from any modality other than text (e.g., image, video), is first transformed into a textual description of the underlying physical process through Keye-VL 1.5 [36]. The generated 4D video V is subsequently captioned using Keye-VL 1.5 to obtain $\hat{T}$, providing a unified textual representation for downstream physical reasoning. **2) Question Answer Generation.** In the second step, we construct a set of diagnostic questions $\{Q_i\}_{i=1}^N$ derived from both the condition prompt C and canonical physical laws. These questions are designed to cover different dimensions of physical understanding, including fundamental physics, physical event reasoning, force and motion, and thermal or phase transitions. For example, one question might ask whether the sidewalls of a

soda can collapse inward once air is removed. Such questions target observable outcomes that must hold if the generated video is physically plausible. **3) Scoring through QA.** In the third step, we input the generated caption $\hat{T}$ along with the question set $\{Q_i\}$ into a large language model, which produces predicted answers $\{\hat{A}_i\}$. These predictions are then compared with the expected physical outcomes $\{A_i^*\}$. For each question, we assign a binary correctness score defined as $s_i = \mathbb{1}(\hat{A}_i = A_i^*)$. And the overall physical realism score is computed as the mean accuracy across all questions: $S_{\text{phy}} = \frac{1}{N} \sum_{i=1}^N s_i$, $S_{\text{phy}} \in [0, 1]$. A higher value of S_{phy} indicates that the generated video adheres more closely to physical reality, while a lower score reflects violations of basic physical constraints, such as the absence of expected deformations, incorrect motion dynamics, or conservation law inconsistencies. More details about prompt and questions can be seen in **supplementary**.

4.2. 4D Condition Alignment

The Condition Alignment evaluation framework measures how well a generated video conforms to its specified con-

dition, ensuring consistency with the intended storyline (event), character motion (motion), attribute and scene. It follows a systematic three-step process—condition input captioning, question generation, and question answering—to quantitatively evaluate alignment quality. **1) Condition Captioning.** The input condition is first converted into a textual description. **2) Question Generation.** We then construct a set of diagnostic questions probing key alignment aspects, including object/character attributes, spatial relations, scene details, motion patterns, and plot elements. All questions are formulated so their answers can be directly inferred from the video. **3) QA-based Scoring.** Given the generated caption $\hat{T}$ and question set $\{Q_i\}_{i=1}^N$, an LLM produces predicted answers $\{\hat{A}_i\}$. These are compared with ground-truth answers $\{A_i^*\}$ derived from the condition. Each question receives a binary correctness score $s_i = \mathbb{1}(\hat{A}_i = A_i^*)$, and the final alignment score is $S_{\text{align}} = \frac{1}{N} \sum_{i=1}^N s_i \in [0, 1]$. A higher score indicates stronger consistency between the generated video and the input condition. More details about prompt and questions can be seen in **supplementary materials**.

4.3. 4D Consistency

To comprehensively evaluate the spatial-temporal stability of generated videos, we assess the **4D Consistency** from three complementary perspectives: **3D Consistency**, **Motion Consistency**, and **Style Consistency**. These three metrics jointly quantify the geometric, dynamic, and perceptual coherence of the generated scene sequences.

3D Consistency. We measure geometric consistency as the reprojection error of 3D points reconstructed by a dense, differentiable SLAM method [10, 29]. To reduce sensitivity to FPS and camera motion, we compute this error on short temporal clips and average over all clips to obtain the final consistency score for both 3D and 4D videos.

Motion Consistency. We evaluate motion consistency with two complementary components: a feature-based optical-flow similarity metric and an MLLM-based judge. For each consecutive frame pair, we estimate optical flow and compare it with the flow induced by the generated motion field. In parallel, to prevent certain videos from exhibiting minor unintended motions, we apply MLLM-as-Judge as described in Sec. 4.2 to perform yes/no question-answering over each video. It evaluates object-level motion rationality (e.g., continuity, interactions, speed and trajectory plausibility), and we use the mean correctness as the motion rationality score. Together, this **hybrid evaluation method** jointly assesses low-level temporal coherence and high-level semantic consistency of motion.

Style Consistency. We assess style consistency [10] by comparing the Gram matrices of deep features between frames. To reduce sensitivity to video length and FPS, we further compute this metric over short temporal clips and

average the results as the final style consistency score. More details about prompt and questions can be seen in **supplementary materials**.

4.4. Perceptual Quality

To further evaluate the perceptual realism of generated videos, we measure four complementary aspects of perceptual quality. **1) Spatial Quality.** We assess the spatial fidelity and technical quality of individual frames using CLIPQA+ [31]. The overall aesthetic appeal of generated frames is evaluated using CLIP-Aesthetic [27], which estimates human aesthetic preferences based on CLIP embeddings. We combine the outputs of these two models to obtain the overall spatial quality score. **2) Temporal Quality.** To evaluate temporal coherence and visual stability across frames, we employ FastVQA [35], a video quality assessment model designed for efficient temporal perceptual evaluation. **3) 3D Texture Quality.** We further measure 3D texture realism by prompting mPLUG-Owl3 [37] to provide ratings on 3D texture fidelity. More details about prompt and questions can be seen in **supplementary materials**.

5. Experiment

5.1. Experiment Details

Evaluated Metrics: Most of the metrics correspond to standalone methodologies. For perceptual quality, our evaluation integrates technical quality and aesthetic quality (as reflected by the a + b components in the table). For motion consistency, we employ a combination of feature-based similarity and MLLM-based QA (corresponding to the a + b components in the table). **Evaluated Models:** For **Image-to-3D**, we select SyncDreamer [22], V3D [7], MotionCtrl [33], ViewCrafter [40], FlexWorld [6]. For **Image-to-4D**, we select CamI2V [43] and Diffusion-as-Shader [14]. For **Text-to-3D**, we select Director3D [20], Text2NeRF [41], Step1X-3D [19], WonderJourney [38]. For **Text-to-4D**, we select 4Dfy [1] and dreamin4D [44]. For **Video-to-4D**, we select ReCamMaster [2], TrajectoryCrafter [39], EX-4D [15], and Vista [13]. Because the 3D model contains no object motion, we omit evaluations of event control, motion control, physical realism, and motion consistency, but since object rotation is presented as a video, we still assess perceptual temporal quality.

Multi-dimensional comparison of 3D/4D generators. Across the five radar plots in Fig. 7, we observe several consistent strengths and weaknesses shared by current 3D and 4D world generative models. On the positive side, most methods already achieve relatively strong scores on *style consistency*, *scene control* and *attribute control*: both text- and image-conditioned 3D models form near-saturated contours on these axes, and the best 4D models largely preserve object identity and global scene composition when moving from static frames to video. Viewpoint consistency is also

Table 3. Leaderboard for different categories of 3D generation models.

Model	Perceptual Quality			Condition-3D Alignment			3D Consistency		Overall
	Spatial	Temporal	3D Texture	Scene	Attribute	Relationship	Viewpoint	Style	
Image-to-3D									
SyncDreameer [22] (Obj)	0.498+0.480	0.460	0.622	0.772	0.864	0.895	0.117	0.943	0.628
V3D [7] (Obj)	0.496+0.439	0.669	0.670	0.788	0.920	0.872	0.470	0.904	0.692
MotionCtrl [33] (Sce)	0.424+0.508	0.738	0.710	0.927	0.926	0.837	0.970	0.888	0.770
Viewcrafter [40] (Sce)	0.497+0.552	0.797	0.715	0.972	0.954	0.863	0.801	0.698	0.761
FlexWorld [6] (Sce)	0.437+0.508	0.745	0.701	0.944	0.953	0.813	0.931	0.780	0.757
Text-to-3D									
Director3D [20]	0.180+0.511	0.601	0.677	0.737	0.813	0.541	0.991	0.992	0.671
Text2NeRF [41]	0.271+0.514	0.754	0.677	0.798	0.805	0.605	0.988	0.878	0.699
Step1x-to-3D [19]	0.455+0.387	0.828	0.596	0.423	0.605	0.490	0.806	0.971	0.618
WonderJourney [38]	0.284+0.449	0.657	0.703	0.821	0.855	0.683	0.774	0.602	0.619

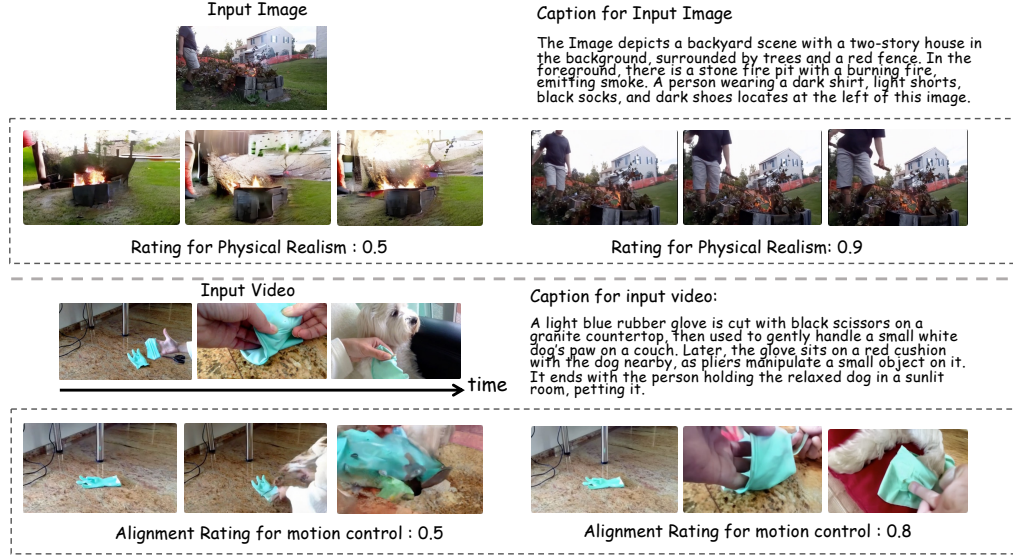


Figure 6. Visualization of some rating examples for physical realism and motion control (4D-Condition Alignment).

Table 4. Performance on Videophy2-test. PC means physical commonsens, SA means semantic adherence, ranging from 0 to 5. The Min(PC,SA) returns the smaller of PC and SA, the Joint is binary (0 or 1), assigned 1 only when both PC and SA larger than 4.

Method	Min(PC,SA) ACC/F1	Joint ACC/F1	Min(PC,SA) PLCC/SRCC	Joint PLCC/SRCC
Ours	0.469/0.439	0.677/0.631	0.408/0.409	0.369/0.376
Videophy2	0.450/0.452	0.770/0.656	0.394/0.385	0.308/0.308
PhygenBench	0.220/0.181	0.360/0.359	0.050/0.066	0.009/0.029
Vbench2	0.220/0.123	0.223/0.187	0.240/0.243	0.225/0.235

reasonably well handled for 3D generation and video-to-4D, suggesting that multi-view geometry priors are effectively exploited by many recent approaches.

In contrast, several dimensions remain challenging across modalities and conditioning types. Motion-related metrics—including *motion alignment*, *dynamics* and *motion consistency*—show a clear gap between the best-performing models and the ideal score, revealing difficulties in generating trajectories that are both semantically correct and physically plausible. For physical realism, such as *optics* and *thermal* cues, remain moderate for 4D generation, suggesting that current architectures struggle to understand physical law under complex viewpoint and illumination

changes. Finally, *event control* and *relationship control* understanding are weak for text-to-4D, highlighting an open problem of grounding rich textual descriptions into coherent interactions over time. Overall, while style and scene-level semantics are largely conquered, temporal reasoning, motion control, and physically realism emerge as common directions for improving future 3D/4D world models.

For detail information in Table 2 and Table 3: **i) Physical Realism:** The rating visualization for physical realism can be seen in Fig. 6. Within each condition, video-based models (e.g., ReCamMaster, TrajectoryCrafter) excel in dynamics and optics, while DiffusionAsShader dominates image-conditioned methods. Text-conditioned models consistently underperform. This suggests that visual inputs, especially image and video, provide much stronger physical grounding than language alone. **ii) 4D Consistency:** Within conditions, DiffusionAsShader outperforms other image-to-4D models, and ReCamMaster achieves the strongest overall consistency among video-based methods. Text-based models remain weaker, with 4Dfy outperforming dreamin4D. This indicates that temporal information is especially beneficial for preserving coherent motion and viewpoint changes

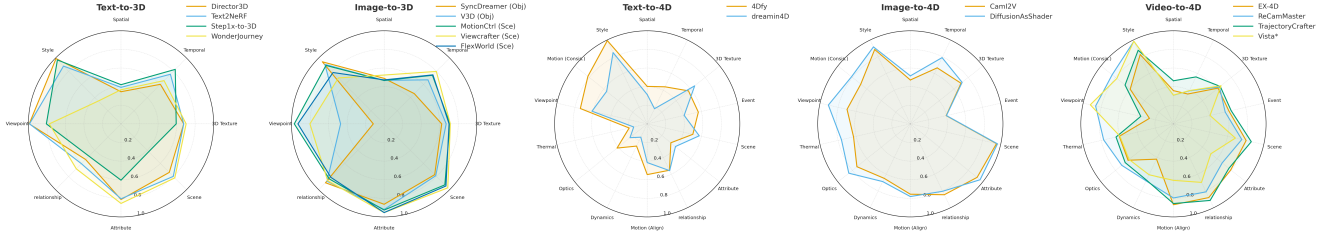


Figure 7. Performance comparison for multiple world generative models.

in 4D generation. **iii) Condition Alignment:** Within conditions, DiffusionAsShader (image) and TrajectoryCrafter (video) achieve the best alignment, while text-based models remain limited. Their results suggest that stronger visual conditioning leads to better control over scene, motion, and attribute preservation. **iv) Perceptual Quality:** Within conditions, DiffusionAsShader achieves the best perceptual quality, followed by TrajectoryCrafter in motion coherence, while text-based models lag behind.

5.2. Ablation for Metric Design

Key-points on Physical Consistency Evaluation. We conducted a subjective evaluation involving around 15 participants on 100 videos to construct a physical-assessment test set, on which we further performed ablation experiments to analyze the metric improvements. The performance comparison of different methods on this subjective test set is presented in Table 5. The results in Table 5 highlight three key findings. First, our LLM-as-judge design consistently outperforms MLLM-based judging, demonstrating the advantage of text-driven reasoning for physical assessment. Second, using more diagnostic questions leads to stronger alignment with human judgment, indicating that richer questioning improves evaluation reliability. Third, adaptive dimension selection (AdaDimen) yields clear gains over fixed dimension settings, showing that LLMs are more effective when allowed to identify semantically relevant physical dimensions rather than relying on predefined ones.

Comparison on Physical Consistency Evaluation. In Table 4, we evaluate our method on the physical-consistency benchmarks covering Physical Commonsense (PC), Semantic Adherence (SA), their conservative minimum of PC & SA, and a binary joint score. Across all evaluation settings, our training-free method achieves performance comparable to—or in some cases exceeding—the trained VideoPhy2 [4] model, while substantially outperforming other training-free baselines. This demonstrates that our approach provides robust and reliable physical reasoning without requiring task-specific training. For more metric ablation studies, please see the **supplementary materials**.

Human Alignment of Other Metrics Design. Table 6 shows that our redesigned metrics achieve better agreement

Table 5. Comparison of judge types, dimension settings, and question counts on Physical Realism Evaluation. “FixDimen” uses pre-defined dimensions; “AdaDimen” applies adaptive dimension.

Setting	Judge (Model)	Dim Type	#Ques	PLCC / SRCC
VBench2	MLLM (Llava-Video)	FixDimen	2	0.341 / 0.379
Ours-v1	LLM(GPT-5→GPT-5)	FixDimen	10	0.351 / 0.402
Ours-v2	LLM (GPT-5→GPT-5)	AdaDimen	10	0.452 / 0.461
Ours	Hybrid(GPT-5→Qwen2.5-VL)	AdaDimen	10	0.247 / 0.237

Table 6. Human correlation (PLCC/SRCC) of different metric designs across alignment and consistency dimensions.

Dimension	Baseline Type	Baseline		Ours	
		PLCC	SRCC	PLCC	SRCC
Attribute Ctrl.	VBench2.0	0.167	0.236	0.483	0.443
Scene Ctrl.	VBench2.0	0.298	0.316	0.502	0.600
Event Ctrl.	VBench2.0	0.320	0.347	0.555	0.580
Motion Ctrl.	VBench2.0	0.250	0.209	0.546	0.564
Rel. Ctrl.	VBench2.0	0.169	0.187	0.458	0.426
Style	WorldScore	0.383	0.457	0.545	0.457
Viewpoint	WorldScore	0.213	0.205	0.543	0.596
Motion	WorldScore	0.332	0.298	0.497	0.450

with human judgments across both 4D-condition alignment and consistency dimensions, yielding more reliable evaluation under different input conditions.

6. Conclusion

In this work, we introduced **4DWorldBench**, a unified, multimodal, and physics-aware benchmark for evaluating next-generation world generation models. Unlike prior video-centric or physics-specific evaluations, 4DWorldBench jointly measures *perceptual quality*, *condition-4D alignment*, *physical realism*, and *4D consistency* within a single, extensible framework. By integrating both MLLM- and LLM-based Question Answering (QA) evaluators, our adaptive methodology enables fine-grained, semantically grounded assessment across diverse text, image, and video conditions. And we also apply hybrid evaluation methods for some dimensions. We hope 4DWorldBench can serve as a standardized platform for benchmarking world generation, promoting fair comparison, deeper understanding, and future innovation toward more coherent, controllable, and physically consistent virtual worlds. In future work, we plan to extend the benchmark to broader real-world scenarios and investigate lightweight evaluation paradigms to further improve scalability and accessibility.

Acknowledgments

This work was supported in part by NSFC under Grant 62371434 and U25B2010, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20252293, the China Postdoctoral Science Foundation-Anhui Joint Support Program under Grant Number 2024T017AH, China Postdoctoral Science Foundation under Grant Number 2025M783529, Anhui Postdoctoral Scientific Research Program Foundation (No.2025A1015), the Fundamental Research Funds for the Central Universities (No. WK2100250064, and ZGCA Project-C20250302).

References

- [1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 5, 6
- [2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14834–14844, 2025. 5, 6
- [3] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *The Thirteenth International Conference on Learning Representations*. 3
- [4] 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. 3, 8
- [5] Xiao Cai, Sitong Su, Jingkuan Song, Pengpeng Zeng, Ji Zhang, Qinrong Du, Mengqi Li, Heng Tao Shen, and Lianli Gao. Gt23d-bench: A comprehensive general text-to-3d generation benchmark. *arXiv preprint arXiv:2412.09997*, 2024. 3
- [6] Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexible-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025. 6, 7
- [7] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, Fuchun Sun, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 6, 7
- [8] Anoop Cherian, Radu Corcodel, Siddharth Jain, and Diego Romeres. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*, 2024. 3
- [9] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023. 3
- [10] 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*, pages 27713–27724, 2025. 1, 2, 3, 6
- [11] Shivam Duggal, Yushi Hu, Oscar Michel, Aniruddha Kembhavi, William T Freeman, Noah A Smith, Ranjay Krishna, Antonio Torralba, Ali Farhadi, and Wei-Chiu Ma. Eval3d: Interpretable and fine-grained evaluation for 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13326–13336, 2025. 3
- [12] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhua Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in conditional video generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4638–4662, 2025. 2
- [13] 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, 2024. 5, 6
- [14] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 5, 6
- [15] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025. 5, 6
- [16] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2
- [17] 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 *International Conference on Machine Learning*, pages 28991–29017. PMLR, 2025. 2
- [18] Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning*, pages 35685–35709. PMLR, 2025. 3
- [19] Weiye Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, et al. Step1x-3d: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. 6, 7
- [20] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation

- from text. *Advances in neural information processing systems*, 37:75125–75151, 2024. 6, 7
- [21] Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816, 2024. 2
- [22] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*. 6, 7
- [23] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tiejong Zeng, Raymond 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*, pages 22139–22149, 2024. 2
- [24] 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 *International Conference on Machine Learning*, pages 43781–43806. PMLR, 2025. 1, 2, 3
- [25] 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*, pages 948–958, 2026. 3
- [26] 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. In *International Conference on Machine Learning*, pages 50338–50362. PMLR, 2025. 1
- [27] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 6
- [28] 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 Computer Vision and Pattern Recognition Conference*, pages 8406–8416, 2025. 2
- [29] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 6
- [30] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3
- [31] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023. 6
- [32] Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13629–13638, 2025. 2
- [33] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 6, 7
- [34] Ziyi Wang, Zhi Gao, Boxuan Yu, Zirui Dai, Yuxiang Song, Qingyuan Lu, Jin Chen, and Xinxiao Wu. Vudg: A dataset for video understanding domain generalization. *arXiv preprint arXiv:2505.24346*, 2025. 3
- [35] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fastvqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pages 538–554. Springer, 2022. 6
- [36] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025. 5
- [37] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multimodal large language models. In *The Thirteenth International Conference on Learning Representations*, 2024. 6
- [38] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snaveley, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 6, 7
- [39] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 100–111, 2025. 5, 6
- [40] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 6, 7
- [41] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 30(12):7749–7762, 2024. 6, 7
- [42] 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. 1, 2, 3
- [43] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video dif-

fusion model. *arXiv preprint arXiv:2410.15957*, 2024. [5](#), [6](#)

- [44] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text- and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7300–7309, 2024. [5](#), [6](#)