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 Yu[1], Xingrui Wang[1], Xinyi Chen[1], Xinge Peng[1], Xin Li[1], 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/.

* Equal contribution †Corresponding author

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*, *autonomous driving*, *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 [44]) 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 [9]) 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. Lastly, physics-aware benchmarks (e.g., PhyGenBench [24]) explicitly focus on testing adherence to physical laws through curated prompts and rule-based evaluation pipelines. Although valuable for physical reasoning, these benchmarks often lack modality diversity, semantic richness, and scal-

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

- **Comprehensive, multimodal benchmark for world generation.** 4DWordBench 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 [13, 18, 44]. 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 [44], 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 [29] and TC-Bench [11] 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 [33] 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 focuses 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 [28]) 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 [10] 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 [9] 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

Table 1. Evaluation dimensions in our 4DWorldBench. Four main dimensions expand into multiple sub-dimensions.

Main Dimension	Sub-dimensions
Perceptual Quality	Spatial Quality; Temporal Quality; 3D Texture
Condition-4D Alignment	Event Control; Scene Control; Attribute Relationship Control; Motion Control
Physical Realism	Dynamics; Optics; Thermal
4D Consistency	Viewpoint Consistency; Motion Consistency; Style Consistency

Table 2. 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 [9]	Text	✓	*partrial*	–	✓
VBench [16]	Text	✓	✓	*partrial*	–
VBench2.0 [44]	Text / Image	✓	✓	*partrial*	–
PhyGenBench [24]	Text	–	–	✓	–
4DWorldBench (ours)	**Text / Image / Video**	✓	✓	✓	✓

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 diservse 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 [17] focuses on real-world free-fall tasks to evaluate the ability of generative models to reproduce physical phenomena like gravity and dynamic collisions. LLMPhy [7] 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 [9], 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. 8 For physical modality, to evaluate physical reasoning and control, we incorporate physics-driven prompts and exemplars following a structured taxonomy (see Fig. 8), 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 [44]. 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 [8]. For physical condition, We first collect videos from *WISA* [32], 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 WideRange4D [37]. 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 cat-

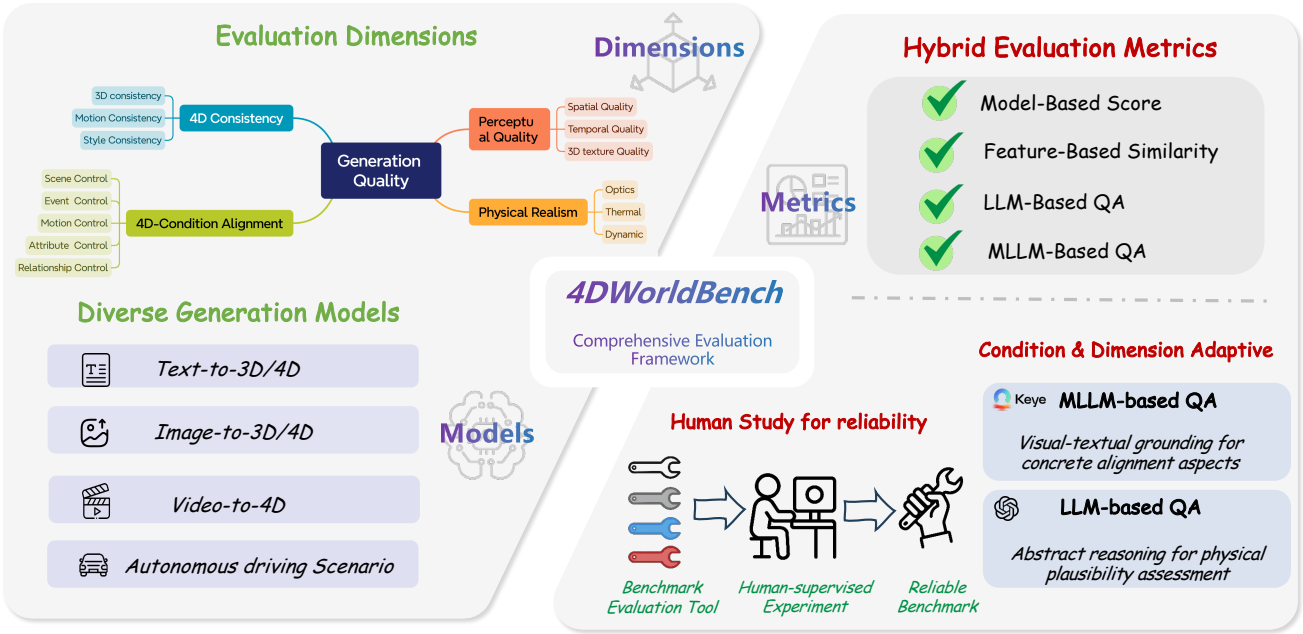

Figure 1. Overview of the 4DWorldBench framework. The benchmark evaluates generation quality across four key dimensions: 4D consistency, condition alignment, perceptual quality, and physical realism. It supports diverse generative settings, including text-, image-, and video-to-3D/4D generation. The 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.

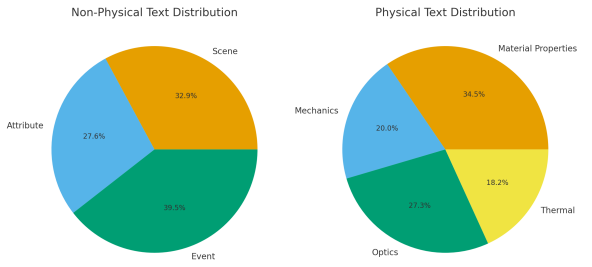

Figure 2. Text Condition Distribution.

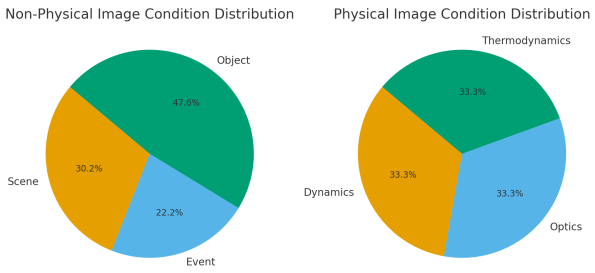

Figure 3. Image Condition Distribution.

egories—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.

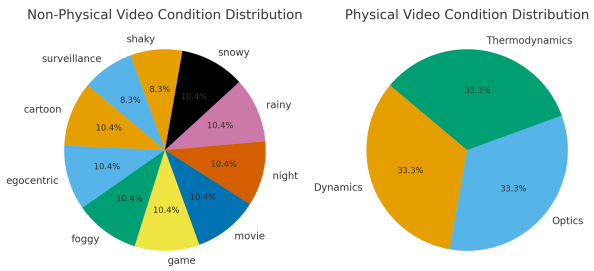

Figure 4. Video Condition Distribution.

For physical condition, we collect videos from *WISA* [32], 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 category, 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 conditioning samples is shown in Fig. 4. Each subset contains 48 samples respectively.

4. Benchmark Metric

Our proposed 4DWorldBench measures 3D/4D generation models across four key dimensions: Physical Realism, Condition-4D Alignment, and 4D Consistency, Perceptual Quality.

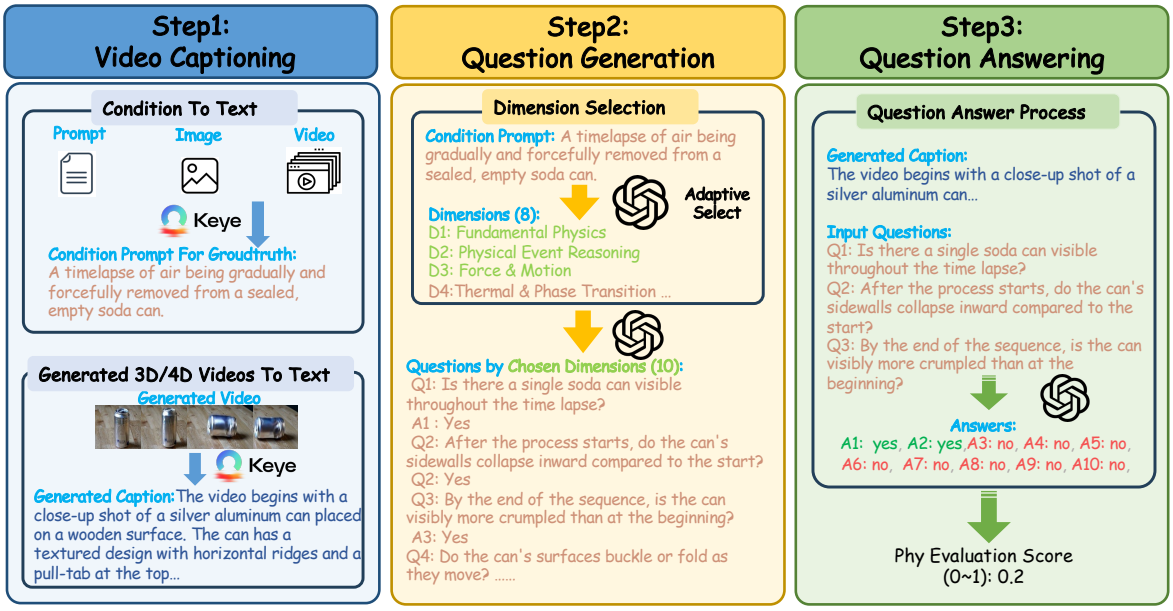

Figure 5. Pipeline of the Physical Realism Evaluation. The framework first unifies text, image, and video conditions into a common textual form, then uses an LLM to adaptively select relevant physical dimensions and generate diagnostic questions. An LLM-based QA module compares predicted and reference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

Table 3. 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															
CamI2V [45]	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 [40]	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 [12]	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 [46]	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

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 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: $S_{\text{phy}} = \frac{1}{N}\sum_{i=1}^{N} s_i, \quad S_{\text{phy}} \in [0, 1]$. A higher value of S_{phy} indicates that the generated video ad-

heres 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 condition, 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 [9, 26, 30]. 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 [9] 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 CLIPIQA+ [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 [38] 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 [6], MotionCtrl [34], ViewCrafter [41], FlexWorld [5]. For **Image-to-4D**, we select CamI2V [45] and Diffusion-as-Shader [14]. For **Text-to-3D**, we select Director3D [20], Text2NeRF [42], Step1X-3D [19], WonderJourney [39]. For **Text-to-4D**, we select 4Dfy [1] and dreamin4D [46] For **Video-to-4D**, we select ReCamMaster [2], TrajectorCrafter [40], EX-4D [15], and Vista [12]. 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.

Table 4. 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									
SyncDreamer [22] (Obj)	0.498+0.480	0.460	0.622	0.772	0.864	0.895	0.117	0.943	0.628
V3D [6] (Obj)	0.496+0.439	0.669	0.670	0.788	0.920	0.872	0.470	0.904	0.692
MotionCtrl [34] (Sce)	0.424+0.508	0.738	0.710	0.927	0.926	0.837	0.970	0.888	0.770
Viewcrafter [41] (Sce)	0.497+0.552	0.797	0.715	0.972	0.954	0.863	0.801	0.698	0.761
FlexWorld [5] (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 [42]	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 [39]	0.284+0.449	0.657	0.703	0.821	0.855	0.683	0.774	0.602	0.619

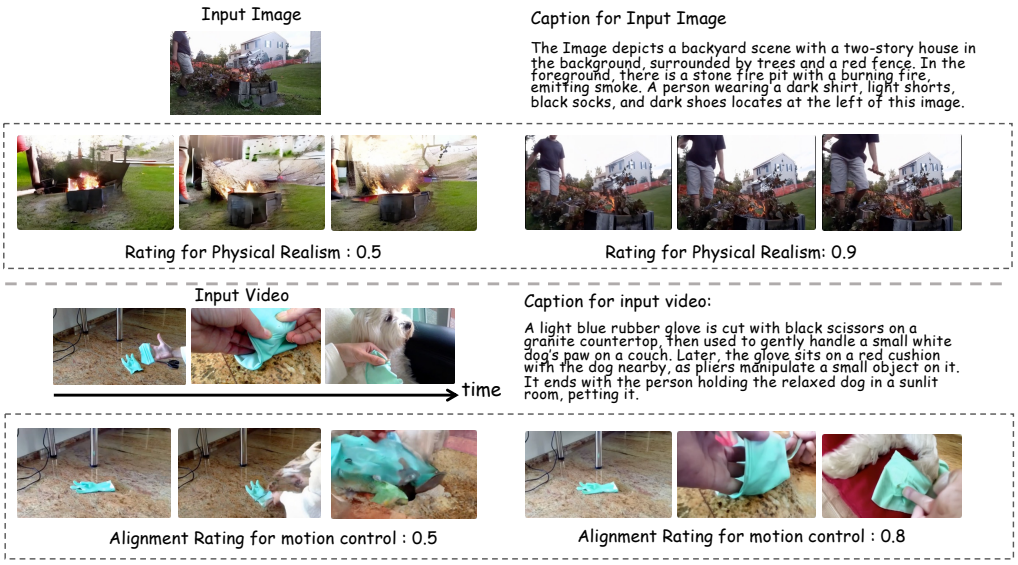

Figure 6. Visualization of some rating examples for physical realism and motion control (4D-Condition Alignment).

Table 5. 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

5.2. Results and Discussion on Physical Realism

Video-conditioned models. Video-to-4D models achieve the excellent performance in physical realism, particularly in modeling dynamic behavior. ReCamMaster [2] and TrajectoryCrafter [40] lead the dynamics sub-metric with scores of (0.680, 0.714, 0.773) and (0.671, 0.667, 0.636), respectively. They also show high optical realism (0.714 and 0.667), benefiting from the spatiotemporal continuity of video inputs. Despite EX-4D [15] having lower dynamics (0.421), it still outperforms text-based models, indicating that even weaker video baselines preserve more physical plausibility than text-conditioned models. **Image-conditioned models.** In Image-to-4D models, Diffusion-AsShader [14] demonstrate superior physical realism on Optics, Thermal and a strong Dynamics score. **It also outperforms other modality-conditioned models**. **Text-conditioned models.** Text-to-4D models exhibit the weakest performance across all physical realism metrics. Even the stronger model, 4Dfy [1], achieves only 0.265 in dynamics and 0.417 in optics, with a low thermal score of 0.200. dreamin4D [46] performs even lower, particularly in optics (0.238), revealing the inherent difficulty of generating physically plausible scenes solely from text. These results emphasize the modality gap in physical realism, with language-conditioned generation still lacking the inductive bias necessary for faithful physical simulation. The rating visualization for physical realism can be seen in Fig. 6.

5.3. Results and Discussion on 4D Consistency

Image-to-3D models: MotionCtrl [34] and FlexWorld [5] achieve high viewpoint consistency (0.970 and 0.931), with strong style retention as well (0.888 and 0.780, respectively). V3D [6] scores moderately (viewpoint: 0.470, style: 0.904), while SyncDreamer [22] performs relatively poorly in viewpoint (0.117) but retains high style (0.943), suggesting potential instability in geometric alignment but

success in style consistency. **Text-to-3D models**: Director3D [20] and Text2NeRF [42] both reach high viewpoint consistency (0.991 and 0.988), indicating strong camera-invariant reconstruction from language. Style-wise, Director3D (0.992) again leads. Step1x-to-3D balances moderate viewpoint (0.806) with good style (0.971), indicating varying trade-offs across architecture. **Text-to-4D models**: 4Dfy [1] consistently outperforms dreamin4D [46] in all three sub-metrics. It achieves especially high performance in motion (0.934+0.705), viewpoint consistency (0.741) and style consistency (0.993), suggesting better viewpoint change and temporal change consistency despite the input modality limitations. **Image-to-4D models**: DiffusionAsShader [14] surpasses CamI2V [45] across all metrics, particularly in motion (0.874+0.750 vs. 0.553+0.816) and viewpoint (0.908 vs. 0.701). Style consistency also favors DiffusionAsShader (0.918 vs. 0.887), indicating better 4D consistency. This suggests that 3D-aware diffusion frameworks may maintain 4D consistency more effectively than 2D video-diffusion methods with geometric priors. **Video-to-4D models**: ReCamMaster [2] achieves the highest overall consistency, with excellent performance in all three areas (viewpoint: 0.862, motion: 0.859+0.834, style: 0.985). In contrast, EX-4D [15] and TrajectoryCrafter [40] show lower viewpoint and motion coherence, despite moderate performance in style. Notably, Vista [12] is for autonomous driving, and while it achieves strong consistency across feature-based metrics (0.942) due to minimal object movement, its performance degrades (0.622) on semantic QA where dynamic object movement is required.

5.4. Results on Condition-3D/4D Alignment

Image-3D models: Viewcrafter [41] achieves the best overall condition alignment, especially in scene (0.972) and attribute alignment (0.954). MotionCtrl [34] and FlexWorld [5] also perform well in attribue and relationship-aligned generation (both large than 0.92). SyncDreamer [22] and V3D [6] show weaker spatial scene understanding and SyncDreamer has lower attribute alignment. **Image-4D models**: DiffusionAsShader [14] outperforms CamI2V [45], achieving higher motion control (0.781), scene control (0.965) and event control (0.4), as well as top attribute alignment (0.963), demonstrating strong fidelity to both static and relational visual cues from input images. **Video-4D models**: TrajectoryCrafter [40] leads in overall condition alignment within video-to-4D models, especially excelling in motion (0.856) and scene (0.757) alignment. Vista has the lowest attribute alignment (0.516), likely due to minimal object movement. EX-4D [15] demonstrates a good balance across different 4D consistency. **Text-3D models**: WonderJourney [39] outperforms other models in scene (0.821) and attribute alignment (0.855), suggesting better adherence to textual semantics. **Text-to-4D models**: 4Dfy [1] and dreamin4D [46] show modest attribute alignment (0.326 and 0.390), consistent with their generation style being limited to object-centric turntable rotations. Text-conditioned models exhibit larger gaps in motion control due to weak temporal grounding. The rating visualization for motion control can be seen in Fig. 6. More visualization examples can be seen in **supplementary materials**.

5.5. Perceptual Quality Analysis

To-3D models: Among 3D generation models, image-conditioned approaches such as Viewcrafter [41] and MotionCtrl [34] achieve the best overall perceptual quality, while text-conditioned ones like Step1x-to-3D [19] exhibit better temporal coherence but weaker spatial fidelity. Overall, image-to-3D models deliver higher spatial and texture quality. **To-4D models:** For 4D generation, DiffusionAsShader (image-conditioned) [14] attains the highest perceptual realism and temporal smoothness, followed by TrajectoryCrafter [40] (video-conditioned) with superior motion coherence, and 4Dfy [1] (text-conditioned) showing modest but improving spatio-temporal stability. Image and video inputs thus favor structure and dynamics.

5.6. 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 6. The results in Table 6 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. 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**.

Table 6. Comparison of judge types, dimension settings, and question counts on Physical Realism Evaluation. "FixDimen" uses predefined 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

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.

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, 7, 8

[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. *arXiv preprint arXiv:2503.11647*, 2025. 5, 6, 7, 8, 17

[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. *arXiv preprint arXiv:2406.03520*, 2024. 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] 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 flexiable-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025. 6, 7, 8, 17

[6] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 6, 7, 8

[7] Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*, 2024. 3

[8] 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

[9] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025. 1, 2, 3, 6, 14, 15, 16, 18

[10] 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. 2

[11] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in conditional video generation. 2

[12] 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, 8

[13] Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, et al. " phyworldbench": A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*, 2025. 2

[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, 7, 8, 17

[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, 7, 8

[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, 3

[17] 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. *arXiv preprint arXiv:2503.09595*, 2025. 3

[18] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025. 2

[19] Weiyu 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, 8

[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. *arXiv:2406.17601*, 2024. 6, 7, 8

[21] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, et al. 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. *arXiv preprint arXiv:2309.03453*, 2023. 6, 7, 8

[23] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 2

[24] 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. 1, 3

[25] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. 3

[26] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 6, 15

[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] Sitong Su, Xiao Cai, Lianli Gao, Pengpeng Zeng, Qinhong Du, Mengqi Li, Heng Tao Shen, and Jingkuan Song. Gt23d-bench: A comprehensive general text-to-3d generation benchmark. *arXiv preprint arXiv:2412.09997*, 2024. 2

[29] 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

[30] 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, 14, 15

[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] Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025. 3, 4

[33] 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

[34] 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, 8, 17

[35] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: 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] Ling Yang, Kaixin Zhu, Juanxi Tian, Bohan Zeng, Mingbao Lin, Hongjuan Pei, Wentao Zhang, and Shuicheng Yan. Widerange4d: Enabling high-quality 4d reconstruction with wide-range movements and scenes. *arXiv preprint arXiv:2503.13435*, 2025. 3

[38] 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 multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 6

[39] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, 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, 8, 17

[40] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 5, 6, 7, 8, 17

[41] 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. *arXiv preprint arXiv:2409.02048*, 2024. 6, 7, 8

[42] 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, 8

[43] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 18

[44] 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, 18

[45] Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. 5, 6, 8, 17

[46] 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, 7, 8

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

Supplementary Material

7. More Details for Benchmark Metrics

7.1. Physical Realism

The prompts for evaluating physical realism in both question–answer generation and caption-based reasoning are presented as follows.

Physics Reasoning QA Template

Video Caption: <Video Caption>

Above is the caption of a video. Please generate yes/no questions for evaluating whether the described scenario follows real-world physics.
Instruction: You are a scientist who designs diagnostic yes/no questions about short real-world scenarios for physics evaluation.
Given: (1) a natural-language description of a real-world scene, and (2) an ordered list of dimensions (e.g., Fundamental Physics, Optics, Material Interaction & Transformation, Force & Motion, Thermal & Phase Transition, etc.).
Returns ten (10) questions in the order provided, with one to four questions per dimension. If there are too few dimensions to ask, more questions per dimension will be required.
Reasoning about reasonable phenomena that should occur in the real world based on the description and posing them as questions, such as a balloon should explode when a sharp object presses into it, water should boil at above 100 degrees, cheese should melt at high temperatures, etc.
You should reason about what should happen in the real world, each question should be crafted to be answerable solely by inspecting the description and focus on visible phenomenon in the description without requiring external knowledge.
If the input description does not contain some problem dimensions or cannot design "yes" answer questions, skip generating questions for that part and move on to the next dimension and do not invent imaginary properties.
Each question must stay strictly within its assigned dimension's scope. Avoid cross-dimension leakage.
Use present tense, neutral tone, and end each question with "(yes or no)".
Return a JSON array of objects, each with:

{ 'dimension': '<dimension name>', 'auxiliary_info': ['<one to four yes/no questions>'] }

Preserve the dimension order. Validate: at most 8 objects, each auxiliary_info has one to four questions, all questions are dimension-appropriate, observable, and answer 'yes'.
Here are some in-context examples:

'questions': [
{ 'dimension': 'Material Interaction & Transformation',
'auxiliary_info': [
'Does the blue and yellow paints visibly exist? (yes or no)',
'Does the blue and yellow paints mix visibly during the stirring process? (yes or no)',
'Does the blue and yellow disappear from the mixed paint? (yes or no)',
'Finally, does the paint become green? (yes or no)'] },

{ 'dimension': 'Force & Motion',
'auxiliary_info': [
'Does the toothpaste tube contact with hands? (yes or no)',
'Does the toothpaste tube deform under stress? (yes or no)',
'Does the toothpaste tube be compressed and the toothpaste be expelled out of the toothpaste tube? (yes or no)'] },

{ 'dimension': 'Thermal & Phase Transition',
'auxiliary_info': [
'Initially, is the river in a liquid state? (yes or no)',
'Finally, does the river freeze? (yes or no)']
}]

System Prompt for Physics Question Generation

Role: JSON-only assistant for physics question generation.

System Prompt:

- "You are an useful assistant that only outputs valid JSON format. Always use double quotes for keys and values, and never use single quotes or any extra text.The format should be: questions:[q1,q2,...]."

Physics Question Design Template

Scene Description: <*Description*>

Above is a natural-language description of a real-world scene and an ordered list of physics-related dimensions. Please design diagnostic yes/no questions for physics evaluation.

Instruction (Physics Reasoning QA Template):

1. You are a scientist who designs diagnostic yes/no questions about short real-world scenarios.
2. Given: (1) a natural-language description of a real-world scene, and (2) an ordered list of dimensions (e.g., Fundamental Physics, Optics, Material Interaction & Transformation, Force & Motion, Thermal & Phase Transition, etc.).
3. Return ten (10) questions in the order provided, with one to four questions per dimension. If there are too few dimensions to ask, more questions per dimension will be required.
4. Reason about reasonable phenomena that should occur in the real world based on the description and pose them as questions (e.g., a balloon should explode when a sharp object presses into it, water should boil above 100 degrees, cheese should melt at high temperatures, etc.).
5. Each question should be crafted to be answerable solely by inspecting the description and focus on visible phenomena in the description, without requiring external knowledge.
6. If the input description does not contain some dimensions or you cannot design a question whose answer is "yes", skip that dimension and do not invent imaginary properties.
7. Each question must stay strictly within its assigned dimension's scope and avoid cross-dimension leakage.
8. Use present tense and neutral tone, and end each question with "(yes or no)".
9. **Output format:** Return a JSON array of objects, each with: { `"dimension": "<dimension name>", "auxiliary_info": ["<one to four yes/no questions>"]` }.
10. Preserve the dimension order. Validate: at most 8 objects, each `auxiliary_info` has one to four questions, all questions are dimension-appropriate, observable, and answer "yes".

In-context examples: The prompt also includes example objects for *Material Interaction & Transformation*, *Force & Motion*, and *Thermal & Phase Transition*, each with several yes/no questions ending with "(yes or no)".

System Prompt for Physics Answer Evaluation

Role: JSON-only assistant for answering physics questions.

System Prompt:

- "You are an assistant that only outputs valid JSON format. Always use double quotes for keys and values, and never use single quotes or any extra text. Example: "answer":"yes" or "answer":"no""

Caption-based Physics Answer Template

Video Caption: <*Caption*>

Question: <*Physics Question*>

Instruction: You are an expert at answering questions based on descriptions of generated videos, which may contain various physically unreasonable. Please answer **yes or no only** for the following question according to the caption. When answering, you should carefully check whether the main objects and behaviors of the question and caption are consistent.

The model must output JSON in the form {`"answer":"yes"`} or {`"answer":"no"`}.

7.2. Condition-4D Alignment

Main Framework For Condition-4D Alignment, it follows a systematic three-step process—condition input captioning, question generation, and question answering—to quantitatively evaluate alignment quality, as illustrated in Fig. 7. Unlike the physical-realism track, which focuses on whether the model can faithfully simulate or reason about complex physical behaviors in 4D space, Condition-4D Alignment targets a different dimension: whether the model can maintain coherent, semantically accurate alignment between user-specified conditions and the generated video content. This dual-track design is motivated by the observation that current Multimodal Large Language Models (MLLM), while proficient at surface-level video question answering, often struggle with complex physical reasoning (e.g., fluid dynamics, force interactions), whereas LLMs excel at abstract reasoning and compositional understanding when provided with high-quality textual descriptions. By decoupling physical realism from condition-driven semantic alignment, our evaluation isolates complementary capabilities and enables a more complete diagnosis of model performance.

Camera Control. Following WorldScore [9], we evaluate camera controllability by comparing the reconstructed camera trajectory with the ground-truth control trajectory. For each generated video, we first estimate frame-wise camera poses using DROID-SLAM [30]. We then measure the angular deviation between the ground-truth and estimated rotations (in degrees) as

$$e_\theta = \arccos\left(\frac{\mathrm{tr}(\mathbf{R}_{\text{gt}}\mathbf{R}^\top) - 1}{2}\right) \cdot \frac{180}{\pi}, \tag{1}$$

and the scale-invariant Euclidean distance between the ground-truth and estimated camera centers as

$$e_t = \left\|\mathbf{t}_{\text{gt}} - s\mathbf{t}\right\|_2, \tag{2}$$

where $\mathbf{R}_{\text{gt}}, \mathbf{R} \in \mathrm{SO}(3)$ are the ground-truth and estimated rotation matrices, $\mathbf{t}_{\text{gt}}, \mathbf{t} \in \mathbb{R}^3$ are the corresponding camera positions, and s is the least-squares scale factor. We combine the rotational and translational errors using the geometric mean to obtain a per-frame camera error

$$e_{\text{camera}} = \sqrt{e_\theta \cdot e_t}. \tag{3}$$

The final camera controllability error for a model is obtained by averaging e_{camera} over all frames of all generated videos, where lower values indicate better adherence to the desired camera trajectory. Note that, unlike other parts of our benchmark, we do not rely on MLLM-based question–answering here, since current multimodal language models exhibit limited ability to accurately reason about fine-grained camera motions.

Prompts for MLLM QA The prompts for evaluating 4D-Condition Alignment as follows.

System Prompt for QA JSON Output

Role: JSON-only assistant for QA generation.

System Prompt:

- "You are an assistant that only outputs valid JSON format. Return the questions as a JSON array of strings."

QA Template for Spatial Relationship Control Evaluation

Please generate detailed yes/no questions about spatial relationships and relative position changes of objects over time based on this caption.

Instruction: You are an expert in caption analysis focusing on object spatial relationship and relative position changes in the whole caption.

Note:

1. Analyze the following video content and generate 5 specific yes/no questions that evaluate the spatial relationship between objects and their relative position changes over time.
2. The description of the video is:
   Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 5 questions.
   3.2. Each question should be answerable with yes/no and the answer of every question should be **yes**.
   3.3. Focus on spatial relationships and relative position changes.
   3.4. Questions should be specific to the video content.
   3.5. The output should be a JSON list of strings.

QA Template for Attribute Control Evaluation

Please generate detailed yes/no questions about dynamic attributes and object transformations.

Instruction: You are an expert in video analysis focusing on dynamic attributes and object transformations.

Note:

1. Analyze the following video content and generate 5 specific yes/no questions that evaluate whether objects in the video show dynamic changes in their attributes (color, size, shape, texture, state, etc.).
2. The description of the video is:
   Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 5 questions.
   3.2. Each question should be answerable with yes/no.
   3.3. Focus on temporal changes and object transformations.
   3.4. Questions should be specific to the video content.
   3.5. The output should be a JSON list of strings.

QA Template for Event Control

Please generate yes/no questions that evaluate the story plot, character actions, and narrative progression in chronological order.

Instruction: You are an expert in story analysis and event understanding.

Note:

1. Analyze the following video content and generate 10 specific yes/no questions that evaluate the event,

story plot, character actions, and narrative progression in chronological order.
2. The description of the video is:
Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 10 questions.
   3.2. Each question should be answerable with yes/no.
   3.3. Questions should follow the chronological order of events.
   3.4. Focus on story elements, character actions, and plot development.
   3.5. Questions should be specific to the video content.
   3.6. The output should be a JSON list of strings.

QA Template for Complex Scene Control

Please generate yes/no questions about detailed scene content in temporal order.

Instruction: You are an expert in scene analysis focusing on complex scene and environments.

Note:
1. Analyze the following video caption and generate 10 specific yes/no questions that evaluate the detailed content of the landscape and environment of the video caption in time order.
2. The description of the video is:
Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 10 questions.
   3.2. Each question should be answerable with yes/no and the answer of every question should be **yes**.
   3.3. Focus on detailed landscape and environment elements.
   3.4. Raise questions about the landscape and scene content of the video caption in time order.
   3.5. The output should be a JSON list of strings.

QA Template for Motion Control

Please generate yes/no questions about the temporal order and sequence of motions.

Instruction: You are an expert in motion analysis and temporal motion sequence understanding.

Note:
1. Analyze the following video caption and generate 10 specific yes/no questions that evaluate the temporal order and sequence of motions described in the video caption.
2. The description of the video is:
Video content: {*content*}
3. Requirements:
   3.1. Generate exactly 10 questions.
   3.2. Each question should be answerable with yes/no and the answer of every question should be **yes**.
   3.3. Focus on temporal order and sequence of motions.
   3.4. Raise specific questions about the existing motions in the video to validate whether the motions in the video are consistent with those described in the caption in time order.
   3.5. The output should be a JSON list of strings.

7.3. 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 using the reprojection error of 3D points reconstructed by a dense, differentiable SLAM pipeline [9, 26, 30]. For each temporal clip $c \in \mathcal{C}$, the clip-level reprojection error is

$$e^{(c)}_{\text{reproj}} = \frac{1}{|\mathcal{V}c|} \sum (i,j) \in \mathcal{V}c \, |\mathbf{p}^* ij - \Pi(\mathbf{P}_{ij})|_2 \,, \qquad (4)$$

where $\mathcal{V}c$ is the set of co-visible pixels inside clip c, $\mathbf{p}^* ij$ is the observed pixel, $\mathbf{P}_{ij}$ is the reconstructed 3D point, and $\Pi(\cdot)$ is the projection operator. The final 3D consistency score is obtained by averaging over all clips:

$$e_{\text{3D}} = 1 - nomarlize(\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} e^{(c)}_{\text{reproj}}). \qquad (5)$$

Motion Consistency. We evaluate motion consistency using a flow-based temporal smoothness metric [9] and an MLLM-based motion rationality score. For each clip c with T_c frames, the flow error is

$$e^{(c)}_{\text{flow}} = \frac{1}{T_c - 1} \sum_{t=1}^{T_c - 1} |\mathbf{F}t \to t+1 - \mathbf{F}'t \to t+1|_2 \,, \quad (6)$$

where $\mathbf{F}t \to t+1$ is the estimated optical flow and $\mathbf{F}'t \to t+1$ is the flow induced by the predicted motion field. In parallel, an MLLM performs video-level yes/no question–answering to assess object-level motion rationality (continuity, interactions, and trajectory plausibility). Let $s^{(c)}_{\text{QA}} \in [0,1]$ denote the mean correctness of its answers for

clip c. The final motion consistency score averages the two components:

$$e_{\text{motion}} = \left(1 - normalize(\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} e_{\text{flow}}^{(c)})\right) + s_{\text{QA}}, \quad (7)$$

where a higher score indicates better temporal coherence and semantic rationality.

Prompt The prompts for evaluating Motion Consistency as follows.

QA Template for Motion Rationality Evaluation in Video Consistency

Please generate yes/no questions that assess the physical plausibility and consistency of object and scene motion.

Instruction: You are an expert in physics and spatiotemporal reasoning, with deep knowledge of real-world motion and dynamics in 4D space (3D + time).

Note:

1. Evaluation Goal: Assess the motion consistency of a 4D generation model, focusing on whether object and scene dynamics evolve plausibly over time.
2. Core Evaluation Principles:
   - 2.1. Temporal Continuity: Are object trajectories and transformations smooth over time, without temporal flickering or abrupt discontinuities?
   - 2.2. Inter-object Interaction: Are interactions (e.g., collisions, pushes, pulls) physically reasonable and temporally aligned?
   - 2.3. Speed and Acceleration Coherence: Are velocity and acceleration patterns consistent with the object's mass, size, and environment?
   - 2.4. Scene-wide Consistency: Do all objects in the scene obey coherent motion logic, including global camera motion if present?
3. Analyze the following video content and generate 5 specific yes/no questions that evaluate these motion rationality principles.
   Video content: {*content*}
4. Requirements:
   - 4.1. Generate exactly 5 questions covering the above principles.
   - 4.2. Each question should be answerable with yes/no.
   - 4.3. Questions must assess physical realism and motion logic.
   - 4.4. Focus on detecting unrealistic physics violations.
   - 4.5. Questions should be specific to the video content.
   - 4.6. The output should be a JSON list of strings.

Style Consistency. We measure style consistency using a VGG-based perceptual metric [9]. For each clip c, we compute the Gram-matrix distance between the first and last frame of the clip:

$$e_{\text{style}}^{(c)} = \left| G(\mathbf{I}^{(c)}1) - G(\mathbf{I}^{(c)}T_c) \right|_F, \quad (8)$$

where $G(\cdot)$ denotes the Gram matrix of deep features. The final score averages clip-level results:

$$e_{\text{style}} = 1 - normalize(\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} e_{\text{style}}^{(c)}). \quad (9)$$

8. More Details for Experiment Metrics

SRCC and PLCC. We quantify the agreement between our benchmark metric and human subjective scores using the Pearson linear correlation coefficient (PLCC) and the Spearman rank-order correlation coefficient (SRCC). PLCC measures the linear correlation between the predicted scores and human ratings, while SRCC evaluates their monotonic relationship. Higher PLCC and SRCC values (closer to 1) indicate that the metric is more consistent with human perception.

9. More Analysis for 3D/4D Generation Models

Camera-control evaluation. As shown in Tab. 7, our method *ReCamMaster* achieves the highest normalized camera-control score. For *TrajCrafter*, we observe that horizontal motions (pan left / pan right) are often entangled with unintended rotational changes of the viewing direction, leading to noticeable drift from the target trajectory and thus a lower camera-control score. In contrast, *FlexWorld* tends to follow the target path reasonably well at the beginning of the sequence, but exhibits off-trajectory rotations in the later part of the video, which also degrades its overall performance.

For all methods, we generate camera trajectories using a set of six canonical motion primitives: *up*, *down*, *pan left*, *pan right*, *zoom in*, and *zoom out*. The superior performance of *ReCamMaster* indicates a better ability to preserve these intended motions without introducing spurious rotations or deviations from the desired camera path.

Multi-dimensional comparison of 3D/4D generators. Across the five radar plots in Fig. 8, 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*

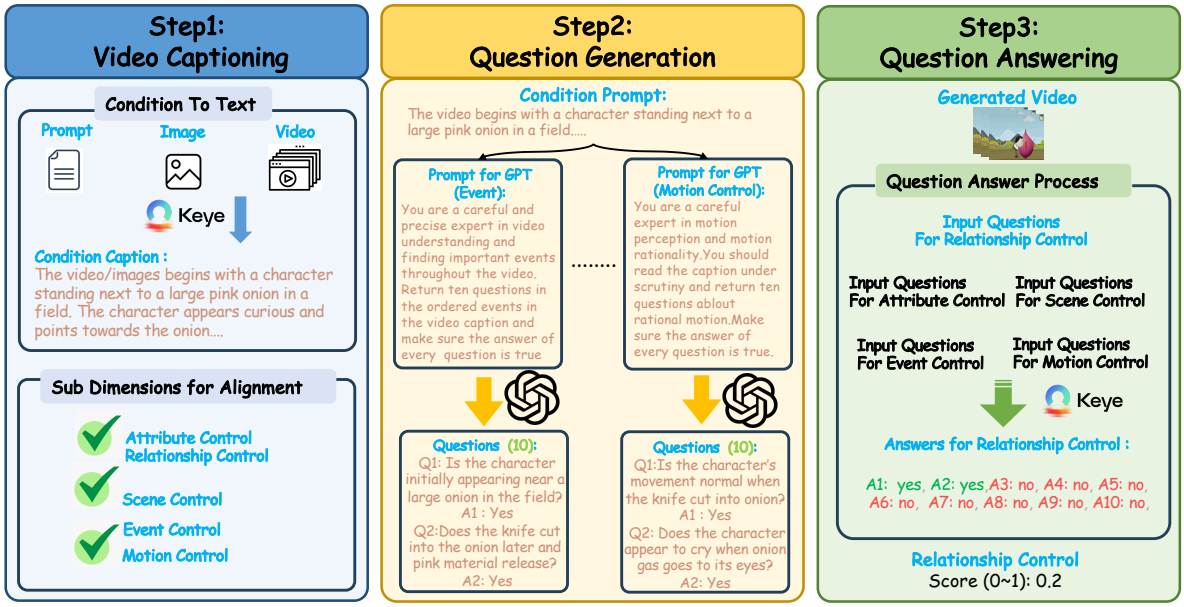

Figure 7. Overall pipeline of 4D-Condition Alignment Evaluation. The framework converts multimodal conditions into text, generates fine-grained and dimension-specific questions for sub-aspects, and uses an MLLM-based QA process to assess the alignment score

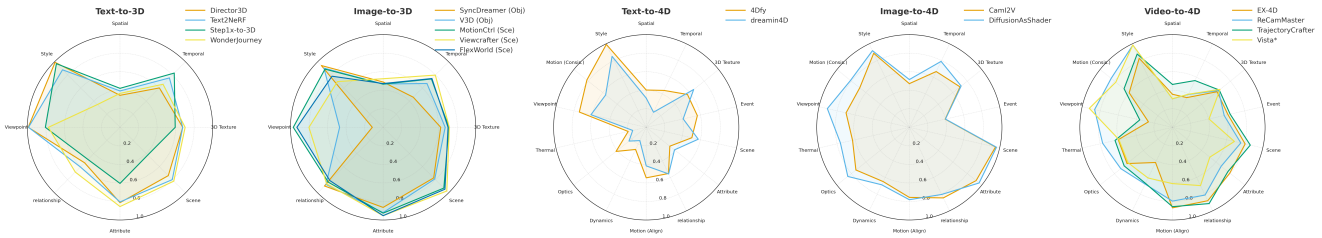

Figure 8. Performance comparison for multiple world generative models.

Method	Score
CamI2V [45]	0.834
DiffusionAsShader [14]	0.793
FlexWorld [5]	0.528
MotionCtrl [34]	0.798
ReCamMaster [2]	0.889
TrajCrafter [40]	0.700
WonderJourney [39]	0.825

Table 7. Camera control scores (higher is better).

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 consisrency is also 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 understanding 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.

10. More ablations for benchmark metrics

We conduct a user study with 10 participants who provide subjective ratings on attribute control and style consistency for 100 generated videos. These ratings are averaged as the ground-turth score for our metric selection.

Table 8. Correlation results on our improved 4D-condition Alignment

Method	Attribute Control PLCC	Attribute Control SRCC
Baseline	0.167	0.236
Ours	0.483	0.443

Table 9. Correlation results on our improved 4D consistency.

Method	Style Consistency PLCC	Style Consistency SRCC
Baseline	0.383	0.457
Ours	0.545	0.457

As summarized in Tab. 8, we compare our redesigned metric against a *Baseline* setting, which use llava-video [43, 44] for question answering. And in Tab. 9, we compare our redesigned metric against a *Baseline* setting, which compute consistency metric at the video-level [9].

4D-Condition Alignment. Tab. 8 demonstrates that our redesigned metric achieves significantly higher correlation with human judgment compared to the previous configuration. For attribute alignment, our method achieves a PLCC of 0.483 and SRCC of 0.443, representing a substantial gain over the previous setting (PLCC: 0.167, SRCC: 0.236). This improvement indicates that our enhanced question–answer formulation and the use of Keye-VL lead to more accurate semantic understanding, resulting in metrics that better reflect human perception.

Style Consistency. On style consistency, our clip-based evaluation approach further improves correlation with human assessments. Specifically, we observe a PLCC increase from 0.383 to 0.545, while the SRCC remains stable (Previous: 0.457, Ours: 0.457). The improved PLCC suggests that our modified video clip-based evaluation yields more stable and reliable style predictions, especially under varying motion and frame rate conditions, despite SRCC saturation. This highlights the benefit of localized, temporal-aware scoring in capturing perceptual appearance stability across generated videos.