WorldScore: A Unified Evaluation Benchmark for World Generation

Haoyi Duan* Hong-Xing Yu* Sirui Chen Li Fei-Fei Jiajun Wu
Stanford University
https://haoyi-duan.github.io/WorldScore/

1. Introduction

Recent advances in visual generation have sparked growing interest in world generation. The rapid progress in video generation [3, 19], 3D scene generation [4, 20, 21], and 4D scene generation [1] has shown generating high-quality individual scenes. However, as the concept of world generation expands, users demand to generate more comprehensive worlds that seamlessly integrate multiple varied scenes with detailed spatial layout controls rather than disconnected individual environments.

Achieving this vision requires a unified evaluation benchmark that systematically assesses different types of world generation models across large-scale, diverse worlds. Existing benchmarks mainly focus on video generation and evaluate only individual scene generation. For example, VBench [10] primarily evaluates text-to-video (T2V) tasks without explicit spatial layout control, restricting their evaluations to single scenes (Figure 1). Moreover, current benchmarks often lack camera specifications and reference images, making them incompatible with many state-of-the-art 3D/4D scene generation methods that require these inputs.

We introduce WorldScore, a unified benchmark for world generation. Our key design is to decompose world generation into a sequence of next-scene generation tasks, where each step is characterized by a triplet of `(current scene, next scene, layout)`. For unified evaluation across different methods, we provide both an image and a text prompt for a `current scene`, as well as both camera matrices and a textual description for a `layout` specification. This design allows our WorldScore benchmark to evaluate various approaches including 3D, 4D, text-to-video, and image-to-video models on large-scale world generation. All methods are evaluated on a common output format, i.e., rendered or generated videos, to enable direct comparison of generation across different types of approaches.

Our evaluation metric, WorldScore, is computed by aggregating three key aspects: *controllability*, which measures the adherence of the generated worlds w.r.t. control inputs; *quality*, which measures the fidelity and consistency; *dynamics*, which measures how much the generated worlds exhibit accurate and stable motions.

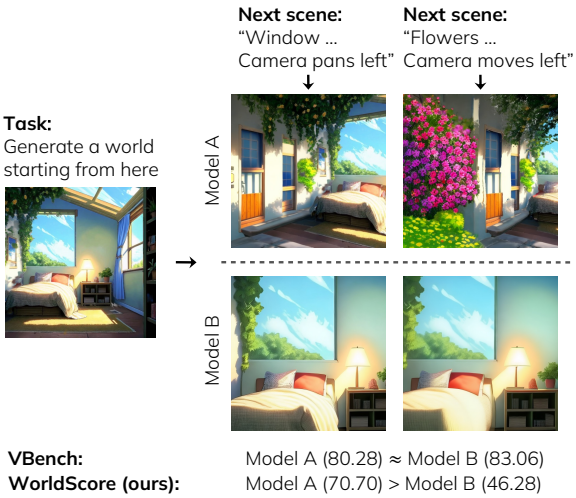

Figure 1. While existing video benchmarks like VBench [10] rate Models A and B similarly based on single-scene video quality, our WorldScore benchmark differentiates their world generation capabilities by identifying that Model B fails to generate a new scene or follow the instructed camera movement.

To enable a comprehensive assessment, we curate a diverse dataset of 3000 high-quality test examples covering both static and dynamic world generation scenarios across different visual domains. We conduct extensive experiments by evaluating 17 diverse models, including image-to-video models (with 2 leading closed-source models), text-to-video models, 3D scene generation models, and a 4D generation model. Through the comprehensive evaluation, we reveal key insights and challenges in current world generation approaches, providing valuable guidance for future research.

2. The WorldScore Benchmark

Design overview. Our goal is to establish an evaluation benchmark for world generation that unifies different methodological approaches. Our WorldScore benchmark

*Equal contribution.

Benchmark	# Examples	Multi-Scene	Unified	Long Seq.	Image Cond.	Multi-Style	Camera Ctrl.	3D Consist.
TC-Bench [6]	150	✗	✗	✗	✓	✗	✗	✗
EvalCrafter [14]	700	✗	✗	✗	✗	✗	✗	✗
VBench [11]	800	✗	✗	✗	✗	✗	✗	✗
T2V-CompBench [16]	700	✗	✗	✗	✗	✗	✗	✗
WorldModelBench [12]	350	✗	✗	✗	✓	✗	✗	✗
WorldScore (Ours)	**3000**	✓	✓	✓	✓	✓	✓	✓

Table 1. **Comparison of Benchmarks.** Our WorldScore benchmark is designed to evaluate various world generation approaches including 3D, 4D, I2V and T2V models. It is designed to generate multiple scenes with varying sequence lengths.

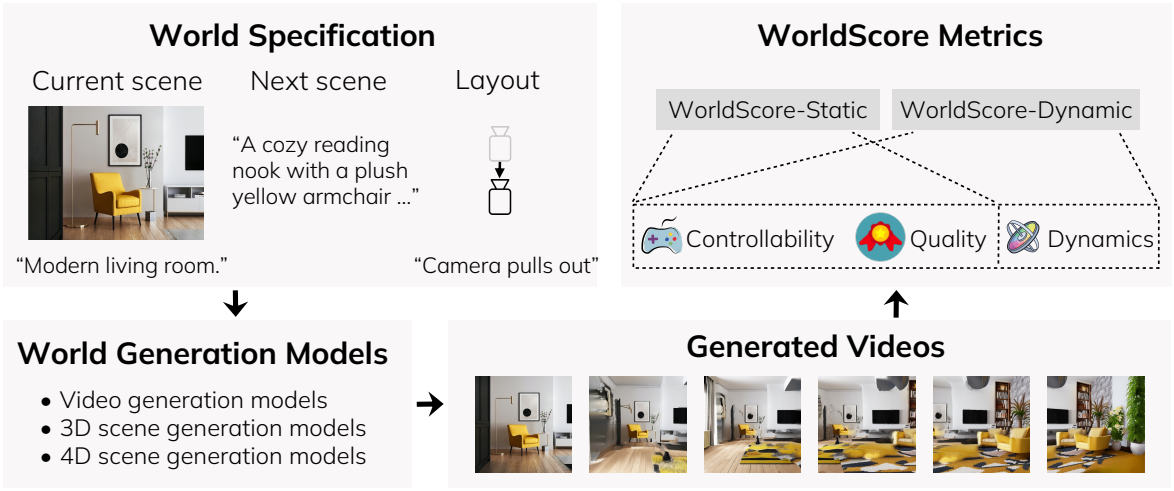

Figure 2. **Overview of the WorldScore benchmark design.** *Top left:* World generation is decomposed into a sequence of next-scene generation tasks, where each step follows a structured world specification defining both spatial layout and semantic content. *Bottom left:* The unified world specification is used to instruct different types of models, including video generation and 3D/4D generation models. *Bottom right:* All models output videos for evaluation. *Top right:* Output videos are evaluated using the WorldScore metrics, which assess three fundamental aspects including controllability, quality, and dynamics.

introduces three key components: (1) a standardized world specification, (2) a carefully curated dataset, and (3) multi-faceted metrics. We show an overview in Figure 2.

2.1. World Specification

Formulation. We decompose the world generation task into a sequence of next-scene generation tasks, where each step is specified by a triplet of $(\mathcal{C}, \mathcal{N}, \mathcal{L})$, where $\mathcal{C} = \{\mathbf{I}, \mathcal{P}\}$ denotes the current scene given by a scene image $\mathbf{I}$ and a text prompt $\mathcal{P}$, $\mathcal{N}$ denotes the next-scene text prompt, and $\mathcal{L} = \{\mathcal{T}, \mathcal{Y}\}$ denotes the layout given by a camera trajectory $\mathcal{T} = (\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_N)$ where $\mathbf{C}_i$ denotes a camera matrix and a text prompt of camera movement $\mathcal{Y}$. Then, a world generation model is instructed to generate a video:

$$\mathbf{V} = g_{\text{world}}(w_{\text{proc}}(\mathcal{C}, \mathcal{N}, \mathcal{L})), \tag{1}$$

where $\mathbf{V}$ denotes a video, g_{world} denotes the world generation model, and w_{proc} denotes a model-specific pre-processing to accommodate the required inputs.

Static and dynamic worlds. We have two types of tasks:

Static world generation: We instruct a model to generate varying-length scene sequences for controllability and quality assessment. Here, the next-scene text prompt $\mathcal{N}$ describes the new scene contents, and the layout $\mathcal{L}$ describes large camera movements.

Dynamic world generation: We instruct a model to generate in-scene motion for dynamics assessment. Here, the next-scene text prompt $\mathcal{N}$ describes the same scene content as $\mathcal{C}$ but with dynamics changes, e.g., an animal moving. The layout $\mathcal{L}$ explicitly specifies a fixed camera position without any camera motion.

2.2. Dataset Curation

Our dataset consists of 3000 examples (world specifications), including 2000 for static world generation and 1000 for dynamic world generation.

Curation on current scene $\mathcal{C}$. For static world generation, we define 10 categories of scenes including 5 indoor and 5 outdoor scene types. Then, we source images from open-source scene datasets and supplement with an online source,

Unsplash. We apply a very rigorous filtering strategy to ensure high quality and high diversity. Then, we query a Vision-Language Model (VLM), GPT-4o, to generate captions $\mathcal{P}$ for these images and do a 10-way classification to put each of them into a category. Finally, we further filter each category by keeping the first 100 highest-quality images, leading to 1000 images $\mathbf{I}$ and their corresponding prompts $\mathcal{P}$.

Then, we create a stylized counterpart for each example in the photorealistic domain. For each example, we randomly pick a style from a set of 7 style candidates, and create a new text prompt $\mathcal{P}$ by adding the style text to the prompt of the photorealistic example. Then, we leverage a commercial style-controlled text-to-image generation model to generate the stylized counterpart image $\mathbf{I}$.

For dynamic world generation, we define 5 categories of motion types and source Unsplash to manually curate 100 images for each of the category. We follow a similar process as in the static world generation examples to create text prompts and stylized counterpart, eventually leading to a total of 1000 examples.

Curation on next-scene text prompts $\mathcal{N}$. Each world generation consists of a sequence of next-scene generation tasks. The next-scene text prompt $\mathcal{N}$ can have varying lengths. To generate coherent and contextually relevant scene sequences, we adopt an auto-regressive scene description generation process [20], that is, we instruct an LLM to generate the next-scene text prompt that should be different from all current scene text prompts.

Curation on layouts $\mathcal{L}$. A layout $\mathcal{L} = \{\mathcal{T}, \mathcal{Y}\}$ is given by a camera trajectory $\mathcal{T} = (\mathbf{C}_1, \mathbf{C}_2, \cdots, \mathbf{C}_N)$ and a text prompt of camera movement $\mathcal{Y}$. We curate a set of 8 camera movements which are widely used in movie industry, including "push in", "pull out", "orbit left", "orbit right", "move left", "move right", "pan left", and "pan right". For each static scene generation example, we randomly assign a layout $\mathcal{L}$ to a next-scene generation task.

2.3. The WorldScore Metrics

Our WorldScore metrics include two overall scores: WorldScore-Static which measures only the static world generation capability, and WorldScore-Dynamic which measures dynamic world generation capability in addition to static worlds. They are defined as the aggregation of several individual metrics in the three key aspects: controllability, quality, and dynamics.

Controllability. We have three metrics, including camera controllability to evaluate how the models adhere to the instructed layout $\mathcal{L} = \{\mathcal{T}, \mathcal{Y}\}$, object controllability to evaluate whether the objects specified in the next-scene prompt $\mathcal{N}$ appear in the generated next scene, and content alignment to assess whether the generated scenes are aligned with the entire text $\mathcal{N}$.

Quality. We have four metrics: (1) 3D consistency evaluates the 3D consistency in the static world videos. This metric focuses on how the geometry of a scene remains stable across frames, regardless of slight changes in visual textures. (2) Photometric consistency: While 3D consistency focuses on geometry while ignoring appearance, photometric consistency measures the stability in appearance (e.g., textures) across frames. (3) Style consistency: We compute the difference between the Gram matrices of the first frame and the last frame of a generated video. (4) Subjective quality: We use automatic metrics to evaluate the human perceptual quality of the generated scenes.

Dynamics. We have three metrics, including motion accuracy to quantify accurate motion placement, motion magnitude to measure a world generation model's ability to create large motions, and motion smoothness.

Score normalization and aggregation. We apply a linear normalization based on empirical bounds to ensure that the final scores fall within the range between zero to one. Then, we compute the arithmetic mean of the dimension scores within control and quality aspects to obtain our **WorldScore-Static**. We incorporate three dynamics dimension scores into the aggregation, resulting in **WorldScore-Dynamic**.

3. Results

We show the WorldScore benchmark results in Table 2. We draw several observations and identify key challenges:

3D models excel in static world generation. From the WorldScore-Static results, we observe that 3D scene generation models generally perform better, e.g., WonderWorld [21] (72.69) and LucidDreamer [4] (70.40) are the top-2, much better than the best video model CogVideoX-I2V [19] (62.15). This is because 3D models inherently have high camera controllability and, thus, better content alignment due to the larger space they can create, as well as high 3D and photometric consistency. However, they do not allow for the generation of dynamic worlds. When extended to 4D for dynamics, 4D-fy [1] does not perform well, likely due to the intrinsic difficulty in 4D scene generation.

Video models lack camera controllability. Even CogVideoX-T2V [19], the best video generation model in camera controllability (40.22), scored much lower than any 3D/4D generation model. This is the main challenge for video generation models to achieve good world generation.

Trade-offs exist in motion smoothness and motion magnitude. Looking at the motion magnitude and motion smoothness metrics, we observe that larger motion often comes at the cost of lower smoothness, revealing current challenge for video models in maintaining both significant motion and natural transitions.

Models	WorldScore		Controllability			Quality				Dynamics		
	-Static	-Dynamic	Camera Ctrl	Object Ctrl	Content Align	3D Consist	Photo Consist	Style Consist	Subjective Qual	Motion Acc	Motion Mag	Motion Smooth
Gen-3 [15]	60.71	57.58	29.47	62.92	50.49	68.31	87.09	62.82	63.85	54.53	27.48	68.87
Hailuo [8]	57.55	56.36	22.39	**69.56**	73.53	67.18	62.82	54.91	52.44	63.46	27.20	70.07
DynamiCrafter [17]	52.09	47.19	25.15	47.36	25.00	72.90	60.95	78.85	54.40	41.11	39.25	26.92
VideoCrafter1-T2V [2]	47.10	43.54	21.61	50.44	60.78	64.86	51.36	38.05	42.63	11.76	**75.00**	18.87
VideoCrafter1-I2V [2]	50.47	47.64	25.46	24.25	35.27	74.42	73.89	65.17	54.85	55.63	25.00	42.49
VideoCrafter2 [2]	52.57	47.49	28.92	39.07	72.46	65.14	61.85	43.79	56.74	47.12	30.40	29.39
T2V-Turbo [13]	45.65	40.20	27.80	30.68	69.14	38.72	34.84	49.65	**68.74**	34.87	40.09	7.48
EasyAnimate [18]	52.85	51.65	26.72	54.50	50.76	67.29	47.35	73.05	50.31	**75.00**	31.16	40.32
CogVideoX-T2V [19]	54.18	48.79	40.22	51.05	68.12	68.81	64.20	42.19	44.67	25.00	47.31	36.28
CogVideoX-I2V [19]	62.15	**59.12**	38.27	40.07	36.73	86.21	88.12	**83.22**	62.44	69.56	26.42	60.15
SceneScape [7]	50.73	35.51	84.99	47.44	28.64	76.54	62.88	21.85	32.75	0.00	0.00	0.00
Text2Room [9]	62.10	43.47	**94.01**	38.93	50.79	88.71	88.36	37.23	36.69	0.00	0.00	0.00
LucidDreamer [4]	70.40	49.28	88.93	41.18	**75.00**	**90.37**	**90.20**	48.10	58.99	0.00	0.00	0.00
WonderJourney [20]	63.75	44.63	84.60	37.10	35.54	80.60	79.03	62.82	66.56	0.00	0.00	0.00
InvisibleStitch [5]	61.12	42.78	93.20	36.51	29.53	88.51	89.19	32.37	58.50	0.00	0.00	0.00
WonderWorld [21]	**72.69**	50.88	92.98	51.76	71.25	86.87	85.56	70.57	49.81	0.00	0.00	0.00
4D-fy [1]	27.98	32.10	69.92	55.09	0.85	35.47	1.59	32.04	0.89	22.22	22.88	**80.06**

Table 2. **WorldScore evaluation results.** The 2nd to 5th sections: closed-source video models, open-source video models, 3D models, 4D models. Abbreviations: Ctrl=Controllability, Align=Alignment, Consist=Consistency, Photo=Photometric, Qual=Quality, Acc=Accuracy, Mag=Magnitude, Smooth=Smoothness.

Larger motion does not necessarily mean more accurate motion placement. The correlation between the motion magnitude and accuracy is weak. This implies that models that can produce large motion do not guarantee correct motion placement to follow instructions. Instead, they could hallucinate unintended camera motion or irrelevant motion. More robust motion modeling may be needed to balance the three dynamics metrics.

References

[1] Bahmani et al. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *CVPR*, 2024. 1, 3, 4

[2] Chen et al. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 4

[3] Chen et al. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 1

[4] Chung et al. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv:2311.13384*, 2023. 1, 3, 4

[5] Engstler et al. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv:2404.19758*, 2024. 4

[6] Feng et al. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv:2406.08656*, 2024. 2

[7] Fridman et al. Scenescape: Text-driven consistent scene generation. *NeurIPS*, 36, 2024. 4

[8] HailuoAI. Hailuo, 2024. https://hailuoai.video/, Accessed: 2025-02-24. 4

[9] Höllein et al. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, 2023. 4

[10] Huang et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 1

[11] Ji et al. T2vbench: Benchmarking temporal dynamics for text-to-video generation. In *CVPR*, 2024. 2

[12] Li et al. Worldmodelbench: Judging video generation models as world models. *arXiv:2502.20694*, 2025. 2

[13] Li et al. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv:2405.18750*, 2024. 4

[14] Liu et al. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, 2024. 2

[15] Runway. Introducing gen-3 alpha: A new frontier for video gneration, 2024. https://runwayml.com/research/introducing-gen-3-alpha, Accessed: 2025-02-24. 4

[16] Sun et al. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. *arXiv:2407.14505*, 2024. 2

[17] Xing et al. Dynamicrafter: Animating open-domain images with video diffusion priors. 2023. 4

[18] Xu et al. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv:2405.18991*, 2024. 4

[19] Yang et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv:2408.06072*, 2024. 1, 3, 4

[20] Yu et al. Wonderjourney: Going from anywhere to everywhere. In *CVPR*, 2024. 1, 3, 4

[21] Yu et al. Wonderworld: Interactive 3d scene generation from a single image. In *CVPR*, 2025. 1, 3, 4