WorldScore: A Unified Evaluation Benchmark for World Generation

Haoyi Duan* Hong-Xing Yu* Sirui Chen Li Fei-Fei Jiajun Wu
Stanford University

Abstract

We introduce the WorldScore benchmark, the first unified benchmark for world generation. We decompose world generation into a sequence of next-scene generation tasks with explicit camera trajectory-based layout specifications, enabling unified evaluation of diverse approaches from 3D and 4D scene generation to video generation models. The WorldScore benchmark encompasses a curated dataset of 3,000 test examples that span diverse worlds: static and dynamic, indoor and outdoor, photorealistic and stylized. The WorldScore metric evaluates generated worlds through three key aspects: controllability, quality, and dynamics. Through extensive evaluation of 20 representative models, including both open-source and closed-source ones, we reveal key insights and challenges for each category of models. Our dataset, evaluation code, and leaderboard can be found at https://haoyi-duan.github.io/WorldScore/.

1. Introduction

Recent advances in visual generation have sparked growing interest in world generation—the creation of large-scale, diverse worlds with various scenes, which finds wide applications in entertainment, education, simulation, and embodied AI. The rapid progress in video generation [1, 6, 10, 88], 3D scene generation [11, 16, 90, 91], and 4D scene generation [3, 85, 89] has shown generating high-quality individual scenes, demonstrating the potential of these models as world generation systems. 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, which is currently absent. Existing benchmarks mainly focus on video generation [15, 45, 46, 48, 92] and evaluate only individual scene generation. For example, VBench [26] primarily evaluates text-to-video (T2V) tasks using curated prompts without explicit spatial layout control, restricting their evaluations to single scenes (Figure 1). Moreover, despite the promising potential of 3D and 4D scene generation methods for world generation, current benchmarks lack essential components such as camera specifications and reference images, making them incompatible with many state-of-the-art 3D/4D scene generation methods that require an image or a camera trajectory as inputs [11, 16, 39, 90, 91].

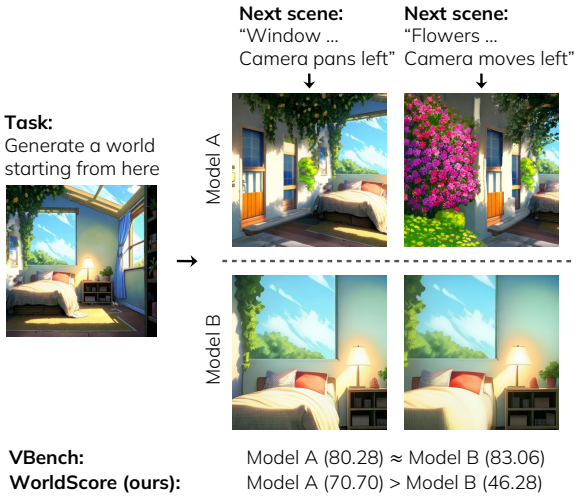

Figure 1. While existing video benchmarks like VBench [26] rate Models A and B similarly based on single-scene video quality, our WorldScore benchmark differentiates their world generation capabilities by identifying that Model B fails to generate a new scene or follow the instructed camera movement. In https://haoyi-duan.github.io/WorldScore/, we show the videos to explain our WorldScore metrics.

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` specifi-

*Equal contribution.

Benchmark	# Examples	Multi-Scene	Unified	Long Seq.	Image Cond.	Multi-Style	Camera Ctrl.	3D Consist.
TC-Bench [15]	150	✗	✗	✗	✓	✗	✗	✗
EvalCrafter [45]	700	✗	✗	✗	✗	✗	✗	✗
FETV [46]	619	✗	✗	✗	✗	✗	✗	✗
VBench [26]	800	✗	✗	✗	✗	✗	✗	✗
T2V-CompBench [71]	700	✗	✗	✗	✗	✗	✗	✗
Meng et al. [48]	160	✗	✗	✗	✗	✗	✗	✗
Wang et al. [78]	423	✗	✗	✓	✗	✗	✗	✗
ChronoMagic-Bench [92]	1649	✗	✗	✗	✗	✗	✗	✗
WorldModelBench [40]	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. Our benchmark also features multiple visual styles, accurate camera control evaluation, and 3D consistency evaluation, all of which are important factors in world generation yet currently missing in existing benchmarks.

cation. 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. Each of these aspects comprises a few distinct metrics, leading to a total of 10 metrics that contribute to computing the WorldScore.

To enable a comprehensive assessment, we curate a diverse dataset covering both static and dynamic world generation scenarios across different visual domains. For static worlds, we include 5 categories of indoor scenes and 5 categories of outdoor scenes with varying sequence lengths. For dynamic worlds, we include 5 distinct types of dynamics such as rigid motion and fluid motion. Additionally, each example in our dataset has a corresponding stylized counterpart sampled from a rich set of candidate styles, allowing the evaluation of various visual domains. In total, our dataset comprises 3000 high-quality test examples that span indoor/outdoor environments and photorealistic/stylized visual domains.

We conduct extensive experiments by evaluating 20 diverse models, including 6 image-to-video models (with 2 leading closed-source models), 7 text-to-video models, 6 3D scene generation models, and a 4D generation model. In summary, our contributions are fourfold:

- We propose the first world generation benchmark, WorldScore, which allows unified evaluation across various approaches including 3D, 4D, I2V, and T2V models.
- We curate a high-quality, diverse dataset for our benchmark evaluation. Our dataset covers diverse static and dynamic scenes across various categories with multiple visual styles.
- We introduce the WorldScore metrics, which aggregate critical aspects in world generation model performances, including controllability, quality, and dynamics.
- Through the comprehensive evaluation of 18 open-source and 2 closed-source models, we reveal key insights and challenges in current world generation approaches, providing valuable guidance for future research.

2. Related Work

Video generation benchmarks. The progress of both open-source [1, 10, 84, 88] and closed-source [2, 6, 20, 58] video generation models has stimulated the proposal of numerous benchmarks [15, 26, 45, 46, 48, 92]. However, most existing benchmarks, such as VBench [26] and WorldModelBench [40], focus on evaluating video generation models based on single-scene video quality without layout control and multi-scene generation. Furthermore, their designs are incompatible with 3D/4D scene generation methods that require camera specification. In contrast, our WorldScore benchmark is designed to focus on evaluating world generation approaches with multi-scene generation tasks, and it is designed to accommodate 3D, 4D, I2V and T2V models. We show a detailed comparison in Table 1.

Video generation models. Recent advances in image generation, including VAEs [36], GANs [5, 18, 30–33, 49], VQ approaches [13, 73], and Diffusion models [23, 52, 68, 70], have fueled explorations in video generation [25, 47, 65, 76, 77]. The advent of Sora [6] has further demonstrated the potential of video models as world generation models [29, 48, 83]. While most recent models focus on text-to-video (T2V) generation [9, 10, 14, 41], developments in image-to-video (I2V) [1, 84, 86, 88, 97] have also been significant. In our WorldScore benchmark, we evaluate both T2V and I2V models as world generation approaches, thanks

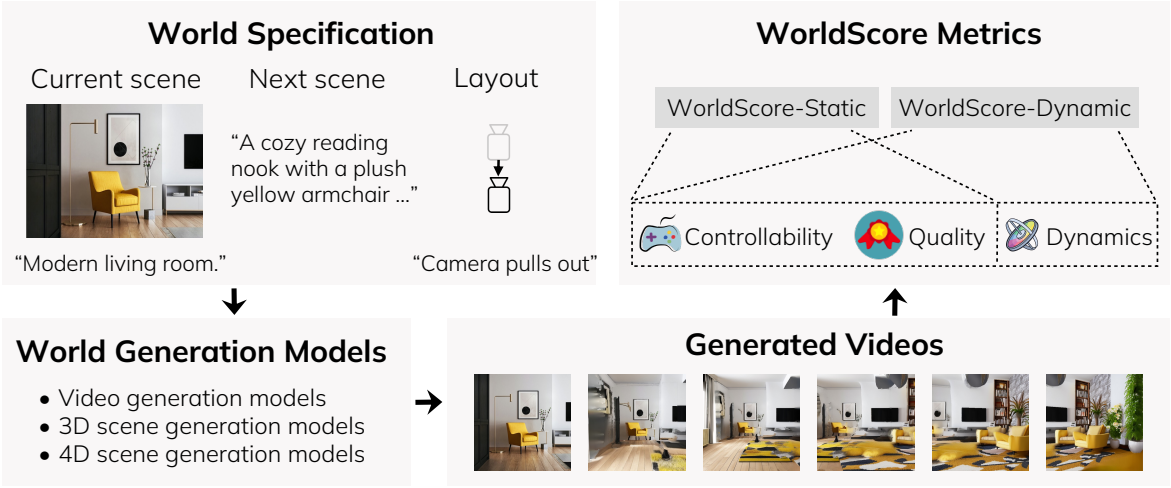

Figure 2. **Overview of the WorldScore benchmark design.** *Top left:* World generation is decomposed into a sequence of next-scene generation tasks, where each step follows a structured world specification defining both spatial layout and semantic content. *Bottom left:* The unified world specification 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.

to our unified design that accommodates both image and text conditioning strategies.

3D scene generation. Besides video models, our WorldScore benchmark also includes 3D and 4D generation methods. Recent 3D scene generation models rely mainly on generative diffusion models [16, 90], which formulate generating scenes in a sequential manner using supervision from 2D image outpainting models. These methods [11, 12, 24, 91] project the synthesized 2D scene extensions into a 3D representation by leveraging depth estimation models [4, 34, 37, 87].

To incorporate dynamics, 4D generation methods [39, 43, 56, 66, 95, 96, 96] further integrate multi-view and video diffusion priors. Due to the difficulty of scene-level generation, most of existing methods focus on object-level generation. Nevertheless, we include 4D-fy [3] in our benchmark due to its open-source accessibility.

3. The WorldScore Benchmark

Design overview. Our goal is to establish an evaluation benchmark for world generation that unifies different methodological approaches. Our WorldScore benchmark 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. We decompose world generation into a sequence of next-scene generation tasks, where each step is defined by a world specification encompassing both spatial layout and semantic content (top left of Figure 2). This world specification enables us to instruct different types of models ranging from 3D/4D scene generation to video generation approaches. The generated outputs, standardized as videos (bottom right of Figure 2), are then evaluated using the WorldScore metrics (top right of Figure 2) that assess three critical aspects: controllability, quality, and dynamics. This unified evaluation approach ensures fair comparison across different methodological paradigms.

3.1. 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 which we detail in Supp. A.

Static and dynamic worlds. We explicitly disentangle the evaluation of dynamics aspect from the controllability and quality aspects due to their distinct natures. To this end, 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.

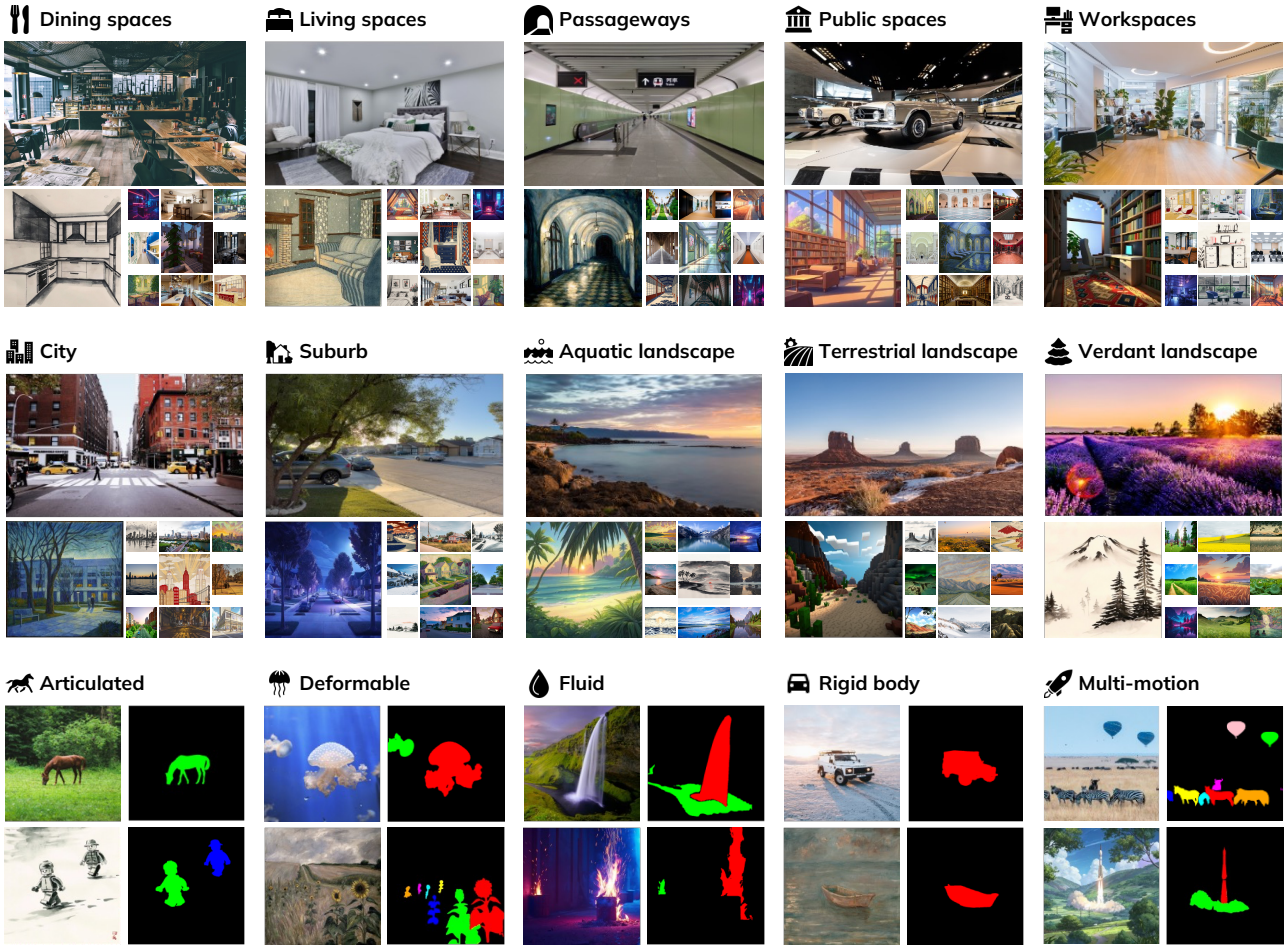

Figure 3. **Showcasing of the current scene images.** *Top two rows:* Static world generation examples are categorized into indoor (first row) and outdoor (second row) scenes, each containing 5 categories. *Bottom row:* Dynamic world generation examples are divided into 5 motion types. Each dynamic example comes with an annotation of motion mask that indicates where the motion should happen.

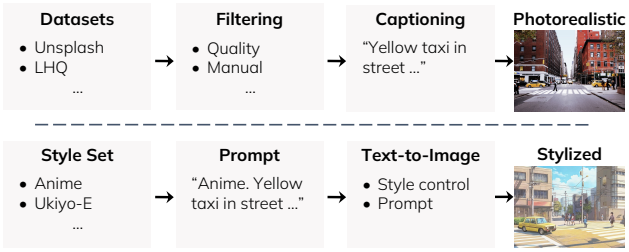

Figure 4. **Curation on the current scene $\mathcal{C}$.** *Top:* Photorealistic worlds. *Bottom:* Stylized counterparts.

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.

3.2. Dataset Curation

Our dataset consists of 3000 examples (world specifications), including 2000 for static world generation and 1000 for dynamic world generation. We show a detailed statistics in Table S4 in the supplementary material.

Curation on current scene $\mathcal{C}$. The current scene $\mathcal{C} = \{\mathbf{I}, \mathcal{P}\}$ is given by an image $\mathbf{I}$ and its text prompt $\mathcal{P}$. We show an illustration of our curation process in Figure 4.

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 [8, 38, 42, 57, 62, 67, 69, 74, 98] and supplement with an online source, Unsplash [7]. We apply a very rigorous filtering strategy to ensure high quality and high diversity (Supp. B.1), leading to approximately 5000 images $\mathbf{I}$ in photorealistic style (they are either real photos or physically-based rendered images). Then, we query a Vision-Language Model (VLM), GPT-4o [51], to generate captions $\mathcal{P}$ for these images and do a 10-way classification to put each of them into a category.

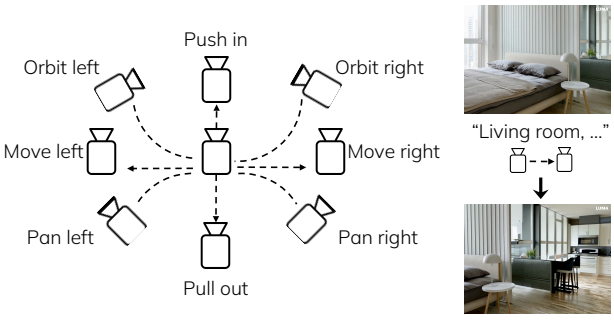

Figure 5. **Curation on layouts $\mathcal{L}$.** *Left:* Camera paths $\mathcal{T}$ and text $\mathcal{Y}$. *Right:* A move-right example.

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 (Supp. B.2). Then, we leverage a commercial style-controlled text-to-image generation model [55] to generate the stylized counterpart image $\mathbf{I}$. We show examples in the top two rows in Figure 3.

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. We show examples in the bottom row in Figure 3.

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. In particular, we consider two cases: (1) a small world where $\mathcal{N}$ consists of only one new scene, and (2) a large world where $\mathcal{N}$ consists of three new scenes.

To generate coherent and contextually relevant scene sequences, we adopt an auto-regressive scene description generation process [90], that is, we instruct an LLM to generate the next-scene text prompt that should be different from all current scene text prompts. For example, for a small world,

$$\mathcal{N} = \text{LLM}(\mathcal{J}, \mathcal{P}), \tag{2}$$

where the LLM takes two inputs: (1) the task specification $\mathcal{J}$ = *"Generate a scene description different from the past scenes."* [1], and (2) a collection of past and current scene descriptions. For a large world which consists of 4 scenes, we repeat this process for 3 times, so that $\mathcal{N} = \mathcal{N}_1 + \mathcal{N}_2 + \mathcal{N}_3$ consists of three individual next-scene prompts. In our generation, 20% of our static world generation examples are large worlds, and the others are small worlds.

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 (left of Figure 5) which are widely used in movie industry. This design achieves two objectives: Firstly, it covers all spatial directions; secondly, it facilitates text-to-video models to take the instruction $\mathcal{Y}$ as most of them are trained on movie clips that often contain these camera movement descriptions. These movements include both intra-scene movements, such as moving into a scene, as well as inter-scene transitions, such as pulling out the camera. For each static scene generation example, we randomly assign a layout $\mathcal{L}$ to a next-scene generation task. We show an example in the right of Figure 5. When the assigned layout is intra-scene, we perform a replacement of $\mathcal{N}$ with $\mathcal{P}$.

We leave details of our dataset curation in Supp. B.

3.3. 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. We briefly introduce each individual metric in the following, and we leave details in Supp. C.

Controllability. We have three metrics.

Camera controllability: To evaluate how the models adhere to the instructed layout $\mathcal{L} = \{\mathcal{T}, \mathcal{Y}\}$, we compute camera errors as follows:

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

where e_θ and e_t are scale-invariant rotation and translation errors with respect to the ground truth trajectory $\mathcal{T}$, respectively. We compute camera errors across all the frames of the generated video $\mathbf{V}$. We leave more details in Supp. C.1.

Object controllability: We evaluate whether the objects specified in the next-scene prompt $\mathcal{N}$ appear in the generated next scene. To this end, we measure the success rate of object detection. Specifically, we leverage a state-of-the-art open-set object detection model [44]. We extract one or two individual object descriptions from the text prompt $\mathcal{N}$. We compute the success rate by matching the detected objects with the object descriptions. This provides a quantitative measure of how well the generated foreground objects adheres to the world specification.

Content alignment: Besides the objects (which typically occupies approximately only $\frac{1}{4}$ of the text prompt length), we also assess whether the generated scenes are aligned with the entire text $\mathcal{N}$ using CLIPScore [22].

[1] This is a brief summary of the actual prompt provided in Supp. B.3.

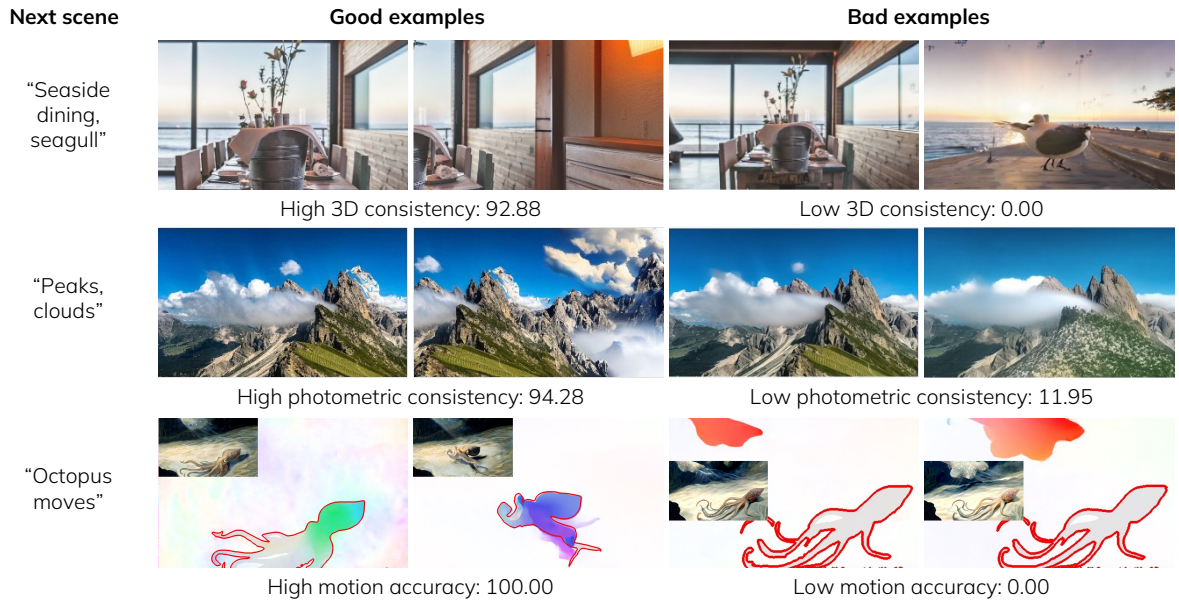

Figure 6. **Typical examples.** *Top: 3D consistency.* The bad example on the right-hand-side has a sudden change in geometry rather than smooth transition. *Middle: Photometric consistency.* The bad example exhibits severe texture shift in the mountain grassland. *Bottom: Motion accuracy.* In the good example, the octopus moves while the jellyfish remains static. For bad example on the right, the jellyfish moves while the octopus remains static. A full version of all metrics is in Figure S3 and Figure S4 in supplementary material. In https://haoyi-duan.github.io/WorldScore/, we show videos to explain our WorldScore metrics.

Quality. We have four metrics.

3D consistency: We evaluate 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. To this end, we use DROID-SLAM [72], a standard SLAM method, to estimate dense pixel-wise depth for each frame, and then we compute the reprojection error between a pair of co-visible pixels in consecutive frames. Since DROID-SLAM is designed to be robust against appearance changes, this metric measures geometric inconsistency. We show an example in Figure 6, and we leave more details in Supp. C.2.

Photometric consistency: While 3D consistency exclusively focuses on geometry, photometric consistency focuses on appearance (e.g., textures). Many video generation models struggle with maintaining consistent object textures, leading to appearance inconsistency issues such as texture flickering. Existing consistency metrics, such as those with CLIP or DINO features [26, 27], focus on categorical identity but fail to capture fine-grained texture changes. For example, the mountain in the middle row of Figure 6 remains a mountain (i.e., the same geometry and semantic class) across frames, but the texture (grass) has been shifted and distorted over time. This cannot be captured by CLIP/DINO features.

To detect photometric artifacts, our photometric consistency metric estimates the optical flow between consecutive frames and computes the Average End-Point Error (AEPE). This metric effectively identifies unstable visual appearance, as shown in Figure 6. We leave more details in Supp. C.3.

Style consistency: We evaluate the style consistency by computing the differences (F-norm) between the Gram matrices [17] of the first frame and the last frame of a single next-scene generation task.

Subjective quality: We use automatic metrics to evaluate the human perceptual quality of the generated scenes. There exists some automatic image assessment metrics [82] and aesthetic metrics [75], and thus we consider ensemble them. To find a combination that best fits human perception, we perform a human study of 400 participants, enumerate different metric combinations, and we pick the combination (CLIP-IQA+ [75] with CLIP Aesthetic [63]) that best matches human preference. We leave more details in Supp. C.4.

Dynamics. We have three metrics.

Motion accuracy: Accurate motion placement is essential in dynamics generation. For example, if a prompt specifies that a car should move while nearby pedestrians remain still, the model should animate the car, not the pedestrians. To quantify this, we introduce motion accuracy, which measures whether the motion specified in the next-scene prompt $\mathcal{N}$ occurs in the designated regions. As shown in the bottom row of Figure 6, the score is calculated by comparing optical flow within the intended region with the flow outside the region. We need to consider the outside flow as it cancels out the global motion caused by unintended camera movements.

Motion magnitude: We measure a world generation model's

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 [58]	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 [20]	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 [84]	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 [9]	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 [9]	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 [9]	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 [41]	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 [86]	52.85	51.65	26.72	54.50	50.76	67.29	47.35	73.05	50.31	75.00	31.16	40.32
Allegro [97]	55.31	51.97	24.84	57.47	51.48	70.50	69.89	65.60	47.41	54.39	40.28	37.81
Vchitect-2.0 [14]	42.28	38.47	26.55	49.54	65.75	41.53	42.30	25.69	44.58	33.59	33.81	21.31
LTX-Video [19]	55.44	56.54	25.06	53.41	39.73	78.41	88.92	53.50	49.08	**76.22**	29.95	71.09
CogVideoX-T2V [88]	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 [88]	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 [16]	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 [24]	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 [11]	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 [90]	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 [12]	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 [91]	**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 [3]	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 of 20 world generation models.** Top: Close-source video models. Middle: Open-source video models. Bottom two rows: 3D and 4D models. Abbreviations: Ctrl=Controllability, Align=Alignment, Consist=Consistency, Photo=Photometric, Qual=Quality, Acc=Accuracy, Mag=Magnitude, Smooth=Smoothness.

ability to create large motions by estimating the optical flow between the consecutive frames of the generated video.

Motion smoothness: Temporal jittering is a common failure mode in dynamic world generation. We utilize a standard video frame interpolation model [93] to generate smooth interpolation as ground truth to evaluate the temporal smoothness of generated videos $\mathbf{V}$. We leave details in Supp. C.7.

Score normalization and aggregation. After computing individual evaluation metrics, we apply a linear normalization and mapping process based on empirical bounds (Supp. C.8) to ensure that the final scores fall within the range between zero to one, and then we scale it by 100. Then, we compute the arithmetic mean of the dimension scores within control and quality aspects to obtain our **WorldScore-Static**. Additionally, we further incorporate three dynamics dimension scores into the aggregation, resulting in **WorldScore-Dynamic**. For 3D scene generation models that do not support dynamic tasks, we assign 0 to each dynamics metric.

4. Results

Validation. We validate our metrics by human study. Our results suggest that WorldScore's metrics align with human preference, and WorldScore is robust to different video resolutions and aspect ratios. We leave details in Supp. D.

Models. We evaluate 20 available world generation models on our WorldScore benchmark. We assess 13 video generation models, including two leading commercial closed-source I2V models—Gen-3 [58] and Hailuo [20], along with 7 well-known open-source I2V models: DynamiCrafter [84], VideoCrafter1-I2V [9], VideoCrafter2 [10], EasyAnimate [86], CogVideoX-I2V [88], LTX-Video [19] and Allegro [97], and 4 open-source T2V models: VideoCrafter1-T2V, T2v-Turbo [41], Vchitect-2.0 [14], and CogVideoX-T2V. Additionally, we evaluate six well-known 3D scene generation models: SceneScape [16], Text2Room [24], LucidDreamer [11], WonderJourney [90], InvisibleStitch [12], and WonderWorld [91]. Moreover, we include an open-source 4D generation model, 4D-fy [3]. We leave details of these models in Table S1 in supplementary material.

4.1. Observations and Challenges

We show the WorldScore benchmark results in Table 2. We identify key challenges in world generation:

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 [91] (72.69) and LucidDreamer [11] (70.40) are the top-2, much better than the best video model CogVideoX-I2V [88] (62.15). This is because 3D models inherently have

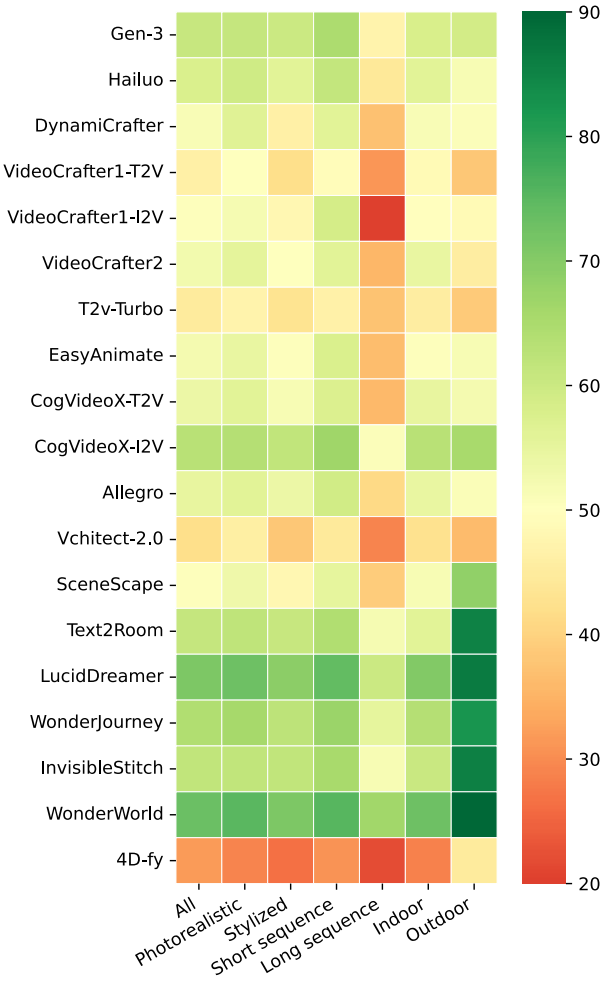

Figure 7. WorldScore-Static across different subdomains.

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 [3] does not perform well, likely due to the intrinsic difficulty in 4D scene generation.

Video models lack camera controllability. Even CogVideoX-T2V [88], 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 static world generation. Recent work in injecting camera conditioning [21, 81] might be a promising solution.

The best open-source video models are as good as closed-source video models. Comparing CogVideoX-I2V [88], with Gen-3 and Hailuo [20], we observe that CogVideoX-I2V scored even higher than both closed-source models in both WorldScore-Static (62.15) and WorldScore-Dynamic (59.12). However, CogVideoX-I2V is not better than them in every aspect. For instance, we observe that CogVideoX-I2V is better at camera controllability yet worse at object controllability and content alignment.

Trade-offs in motion smoothness and magnitude. Comparing motion smoothness and motion magnitude metrics for each method, 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.

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.

Video models are weak in long sequence generation and in outdoor scenes. We further evaluate model performance across different subdomains, and we show WorldScore-Static results in Figure 7. We observe that video generation models struggle significantly with long-sequence (large world generation) tasks. In addition, video models are significantly weaker than 3D models in outdoor scenes, while the gap is smaller in indoor scenes.

T2V models are easier to steer than I2V models. Compare T2V models to I2V models, e.g., CogVideoX-T2V and CogVideoX-I2V, we observe that T2V models generally have higher scores in the controllability aspect and larger motion magnitude, while I2V models have higher scores in quality aspect. Through empirical examination, we find that this is because T2V models are willing to generate larger camera motion, while I2V models tend to stick to the input image viewpoint. This reveals a challenging in controlling I2V models to generate new scene contents. We leave further visualizations in Supp. E.

5. Conclusion

The WorldScore benchmark reveals current limitations in world generation approaches. For 3D models, while they excel in static world generation, extending them to 4D representations and incorporating dynamics remains challenging. For video models, the main challenges include controllability, long-sequence generation, and generating outdoor scenes. These insights point to directions for future research: bridging the gap between 3D and 4D representations, developing more robust controllability mechanisms, and designing architectures capable of handling extended scene sequences. We believe the WorldScore benchmark will serve as a valuable tool for measuring progress toward more capable and versatile world generation systems.

Acknowledgments. This work is in part supported by ONR YIP N00014-24-1-2117, ONR MURI N00014-22-1-2740, NSF RI #2211258 and #2338203, and the Okawa Foundation. We thank Mohamed El Banani and Christoph Lassner for their helpful discussion.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1, 2

[2] Luma AI. Luma dream machine: New freedoms of imagination, 2025. https://lumalabs.ai/dream-machine, Accessed: 2025-02-24. 2

[3] 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. 1, 3, 7, 8, S2

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[5] Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 2

[7] Supported by SQUARESPACE. Unsplash, 2013. https://unsplash.com, Accessed: 2025-02-23. 4, S1

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 4, S1, S4

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 2, 7, S2

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 1, 2, 7, S2

[11] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 1, 3, 7, S2

[12] Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv preprint arXiv:2404.19758*, 2024. 3, 7, S2

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[14] Weichen Fan, Chenyang Si, Junhao Song, Zhenyu Yang, Yinan He, Long Zhuo, Ziqi Huang, Ziyue Dong, Jingwen He, Dongwei Pan, et al. Vchitect-2.0: Parallel transformer for scaling up video diffusion models. *arXiv preprint arXiv:2501.08453*, 2025. 2, 7, S1

[15] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024. 1, 2

[16] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3, 7, S2

[17] Leon A Gatys. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 6

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[19] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 7, S2

[20] HailuoAI. Hailuo, 2024. https://hailuoai.video/, Accessed: 2025-02-24. 2, 7, 8, S1, S2

[21] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 8

[22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 5

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[24] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 3, 7, S2

[25] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[26] 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. 1, 2, 6

[27] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 6

[28] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17307–17316, 2023. S1

[29] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 2

[30] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[31] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[33] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 2

[34] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 3

[35] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. S5

[36] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[37] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 3

[38] Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, and Theo Gevers. Eden: Multimodal synthetic dataset of enclosed garden scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1579–1589, 2021. 4, S1, S4

[39] Yao-Chih Lee, Yi-Ting Chen, Andrew Wang, Ting-Hsuan Liao, Brandon Y Feng, and Jia-Bin Huang. Vividdream: Generating 3d scene with ambient dynamics. *arXiv preprint arXiv:2405.20334*, 2024. 1, 3

[40] 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

[41] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024. 2, 7, S1, S2

[42] Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 473–488. Springer, 2020. 4, S4

[43] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8576–8588, 2024. 3

[44] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025. 5

[45] 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. 1, 2

[46] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[47] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 2

[48] 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, 2

[49] Mehdi Mirza. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[50] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. S6

[51] OpenAI. Hello gpt-4o, 2024. https://openai.com/index/hello-gpt-4o/, Accessed: 2025-02-23. 4, S1

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. S1

[54] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. S5

[55] Recraft. Recraft image generation and editing api, 2025. https://www.recraft.ai/docs, Accessed: 2025-02-25. 5, S1

[56] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 3

[57] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 4, S1, S4

[58] Runway. Introducing gen-3 alpha: A new frontier for video gneration, 2024. https://runwayml.com/research/introducing-gen-3-alpha, Accessed: 2025-02-24. 2, 7, S1, S2

[59] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. S1

[60] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. S4

[61] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. S4

[62] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 4, S1, S4

[63] Christoph Schuhmann. Clip+ mlp aesthetic score predictor. *Clip+ mlp aesthetic score predictor*, 2022. 6, S1, S5, S7

[64] SDXL. 106 styles for stable diffusion xl model, 2023. S1

[65] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[66] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 3

[67] Ivan Skorokhodov, Grigorii Sotnikov, and Mohamed Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14144–14153, 2021. 4, S1, S4

[68] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[69] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 4, S1, S4

[70] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[71] 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. *arXiv preprint arXiv:2407.14505*, 2024. 2

[72] 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, S4

[73] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[74] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 4, S1, S4

[75] 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*, 2023. 6, S1, S5, S7

[76] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[77] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, pages 1–20, 2024. 2

[78] 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. *arXiv preprint arXiv:2412.16211*, 2024. 2

[79] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. S5

[80] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to

structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. S6

[81] 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. 8

[82] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 6, S5

[83] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 2

[84] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. 2023. 2, 7, S2

[85] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024. 1

[86] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024. 2, 7, S1, S2

[87] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 3

[88] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 2, 7, 8, S2

[89] Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *arXiv preprint arXiv:2406.07472*, 2024. 1

[90] 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*, 2024. 1, 3, 5, 7, S1, S2

[91] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 3, 7, S1, S2

[92] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *arXiv preprint arXiv:2406.18522*, 2024. 1, 2

[93] Guozhen Zhang, Chunxu Liu, Yutao Cui, Xiaotong Zhao, Kai Ma, and Limin Wang. Vfimamba: Video frame interpolation with state space models. *arXiv preprint arXiv:2407.02315*, 2024. 7, S5, S6

[94] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. S6

[95] Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023. 3

[96] 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. 3

[97] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024. 2, 7, S2

[98] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 4, S1, S4