# World-R1: Reinforcing 3D Constraints for Text-to-Video Generation

Weijie Wang[1,2,*†] Xiaoxuan He[1,*] Youping Gu[1,*] Yifan Yang[2,‡]
Zeyu Zhang[3] Yefei He[1] Yanbo Ding[2] Xirui Hu[3]
Donny Y. Chen[3] Zhiyuan He[2] Yuqing Yang[2,‡] Bohan Zhuang[1,‡]

[1] Zhejiang University [2] Microsoft Research [3] Independent Researcher

arXiv:2604.24764v4 [cs.CV] 26 May 2026

Recent video foundation models demonstrate impressive visual synthesis but frequently suffer from geometric inconsistencies. While existing methods attempt to inject 3D priors via architectural modifications, they often incur high computational costs and limit scalability. We propose World-R1, a framework that aligns video generation with 3D constraints through reinforcement learning. To facilitate this alignment, we introduce a specialized pure text dataset tailored for world simulation. Utilizing Flow-GRPO, we optimize the model using feedback from pre-trained 3D foundation models and vision-language models to enforce structural coherence without altering the underlying architecture. We further employ a periodic decoupled training strategy to balance rigid geometric consistency with dynamic scene fluidity. Extensive evaluations reveal that our approach significantly enhances 3D consistency while preserving the original visual quality of the foundation model, effectively bridging the gap between video generation and scalable world simulation.

**Project Page:** https://aka.ms/world-r1
**Correspondence:** {yifanyang, yuqyang}@microsoft.com, bohan.zhuang@zju.edu.cn
**Conference:** The $43^{rd}$ International Conference on Machine Learning

* *Equal contribution.* † *Work was done during an internship in MSRA.* ‡ *Corresponding authors.*

## 1 Introduction

Recent advances in visual generation have ignited a paradigm shift, moving beyond simple content creation toward the ambitious goal of world generation. Through extensive pre-training on internet-scale data, video generation models [1–6] are increasingly recognized as precursors to general-purpose world models. These foundation models demonstrate exceptional capabilities in synthesizing high-fidelity visual environments, holding transformative potential for diverse fields such as autonomous driving, robotics, and immersive content creation.

However, despite their impressive visual proficiency, current video foundation models [1, 3] fundamentally focus on image-space generation. They lack an intrinsic understanding of the 3D geometry that governs the real world. Consequently, while they excel at generating short, static-camera clips, they frequently suffer from geometric hallucinations and temporal inconsistencies when synthesizing large camera movements or long-horizon driving scenes. Without explicit 3D constraints, objects may morph, vanish, or distort unphysically, revealing that these models are merely mimicking surface-level correlations rather than simulating a coherent real world.

To address these limitations, recent works [7–15] have attempted to bridge video generation with 3D representations. Some approaches [8, 9, 11] explicitly inject 3D priors into the inference

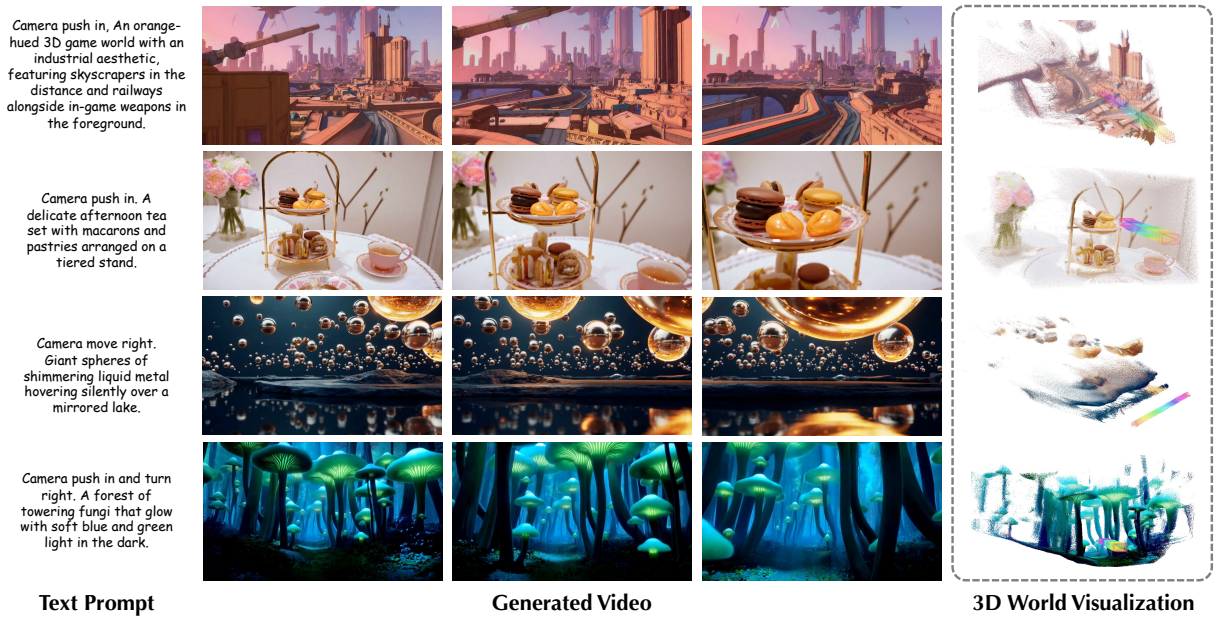


**Figure 1** World-R1 enables geometrically consistent world modeling in video foundation models via reinforcement learning. It accepts text input and generates high-quality 3D consistent videos. We visualize the reconstruction results alongside the video frames. Video results are in our supplementary material.

process of video models, forcing the generation to adhere to static 3D constraints. While these methods improve consistency to some extent, they often incur prohibitive inference costs and restrict generation precision and generalization scope of the model. Such heavy, inference-time constraints prevent these systems from scaling efficiently or maintaining the fluid diversity of the original video models.

Building upon the finding that video foundation models already inherently encode rich 3D geometric information [16, 17], we identify that a path bridging the gap between video generation and world modeling lies in eliciting this latent knowledge rather than merely scaling data or imposing rigid inference constraints. In this paper, we introduce World-R1, a novel framework that injects world-modeling capabilities into video models via reinforcement learning (RL) [18]. Uniquely, our approach achieves this without relying on expensive 3D assets for supervised training, and crucially, without altering the model architecture or inference process.

We leverage the inherent consistency requirements of the real world to construct a robust reward mechanism via analysis-by-synthesis. Specifically, we utilize pre-trained 3D foundation models [19–24] to enforce rigid geometric fidelity through reconstruction and trajectory alignment, while employing vision-language models (VLMs) [25–28] as semantic critics to evaluate the plausibility of rendered novel views (meta-views). To enable precise motion control without architectural modification, we introduce an implicit camera conditioning strategy that embeds trajectory priors directly into the latent noise. Furthermore, we construct a synthetic pure text dataset to dissociate physical learning from visual bias and adopt a periodic decoupled training strategy, which mitigates the suppression of non-rigid dynamics often caused by strict 3D constraints. Driving this alignment via Flow-GRPO [29], our framework empowers existing video generation models to internalize geometric laws, effectively transforming them from 2D frame predictors into geometrically consistent world simulators.

To comprehensively evaluate the capabilities of World-R1, we evaluate both scene generation quality and general video generation performance. Experiments demonstrate that our fine-tuned models significantly improve geometric consistency, achieving an improvement of 10.23dB and 7.91dB on PSNR respectively, while maintaining high scores on general video benchmarks.

Additional analyses in the appendix further validate the gains with a reconstruction-independent multi-view consistency metric, dataset-scaling results, long-video evaluation, scene-complexity breakdowns, and comparisons to 3D-aware generation methods. This indicates that our method effectively injects world modeling capability without compromising original generative diversity or visual quality of the model. Our contributions can be summarized as follows:

- We propose World-R1, a novel paradigm that utilizes RL to align video generation models with 3D foundation models, eliminating the need for cultivating large-scale 3D datasets or computationally expensive inference-time constraints.
- We design a comprehensive reward system integrating pre-trained 3D foundation models and VLMs, optimized via Flow-GRPO [29], allowing the video model to internalize geometric consistency through discriminative feedback.
- We constructed a pure text dataset featuring multi-class, multi-level camera control for the post-training of video generation models.
- Extensive experiments demonstrate that our framework significantly enhances the world-modeling capabilities of video models, yielding superior 3D reconstruction consistency while preserving high-quality visual generation performance.

## 2 Related Works

**Controllable Video Generation.** The emergence of large-scale video foundation models has revolutionized the field of video generation. While early approaches relied on U-Net-based diffusion architectures, recent state-of-the-art models [1–3] have shifted towards DiT architectures. Trained on internet-scale video datasets, these models demonstrate exceptional capabilities in synthesizing high-fidelity visual content. Despite their success, precise camera control within these foundation models remains a persistent challenge. Several methods [30–39] attempt to address this by training auxiliary control modules to introduce explicit camera pose conditioning. However, these approaches primarily focus on trajectory adherence and require additional inputs. Crucially, they fail to guarantee 3D geometric consistency, often resulting in object distortion during complex camera movements. Recognizing that foundation models are inherently insensitive to camera motion without guidance, our approach builds upon the Go-With-The-Flow [40] paradigm. We adopt a module-free strategy for camera control, avoiding the introduction of extra architectural components. Instead, we enhance the instruction following capabilities and 3D consistency through post-training optimization.

**3D-Aware Video Generation.** To mitigate the lack of spatial awareness and 3D consistency in standard video models, recent research has sought to bridge image-to-video(I2V) generation with 3D representations. Several methods [8, 9, 12, 41, 42] attempt to solve this by integrating explicit 3D representations directly into the video generation pipeline. Similarly, Fantasyworld [43] employs a multi-task learning framework by appending a 3D decoder to the video encoder to generate pointmaps. However, these methods typically necessitate significant architectural modifications and incur high inference latency due to the additional computational burden of 3D modules and can only be applied to I2V tasks. Furthermore, their reliance on static 3D aware datasets [44, 45] for training often severely limits the diversity and dynamic adaptability of the output. In contrast, World-R1 avoids architectural modifications entirely and can handle text-to-video (T2V) tasks. Rather than enforcing 3D constraints via external modules, we leverage reinforcement learning to elicit the latent spatial awareness. This allows the model to learn geometry and maintain consistency without explicit 3D guided inference.

**Visual Reinforcement Learning.** RL [18] has recently shown immense potential in aligning generative models with human preferences. While PPO [46] has been the standard for language

models, it incurs prohibitive computational costs when applied to high-dimensional visual data. GRPO [47] has emerged as a more efficient alternative by eliminating the need for a critic network. Building on this, Flow-GRPO [29] adapts the GRPO framework to flow-matching-based generative models. To further accelerate training, Flow-GRPO-Fast [29] injects noise into the deterministic ODE trajectory at randomly selected intermediate steps, switching to SDE sampling. We extend this framework to the 3D consistent video generation by designing tailored reward mechanisms that specifically penalize geometric inconsistencies. This effectively transforms a general-purpose video generator into a geometrically consistent world simulator.

## 3 Preliminaries

Flow-GRPO [29] enhances flow matching models by integrating online RL to optimize generation quality. While standard flow matching relies on deterministic ODE solvers, RL requires stochasticity for effective exploration and advantage estimation. Flow-GRPO addresses this by reformulating the sampling process and applying GRPO.

**Stochastic Sampling via SDE.** To introduce the necessary stochasticity for RL while preserving the marginal distribution of the pre-trained flow model, Flow-GRPO converts the deterministic flow ODE ($\mathrm{d}\mathbf{x}_t = \mathbf{v}_t \mathrm{d}t$) into a reverse-time Stochastic Differential Equation (SDE):

$$\mathrm{d}\mathbf{x}_t = \left[\mathbf{v}_t(\mathbf{x}_t) + \frac{\sigma_t^2}{2t}\left(\mathbf{x}_t + (1-t)\mathbf{v}_t(\mathbf{x}_t)\right)\right]\mathrm{d}t + \sigma_t \mathrm{d}\mathbf{w}, \tag{1}$$

where $\sigma_t$ controls the noise level and $\mathbf{w}$ denotes the Wiener process. Discretizing Eq. (1) yields the stochastic update rule that serves as the policy $\pi_\theta$:

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \left[\mathbf{v}_\theta(\mathbf{x}_t, t) + \frac{\sigma_t^2}{2t}\left(\mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t)\right)\right]\Delta t + \sigma_t\sqrt{\Delta t}\,\epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. This formulation transforms the deterministic sampler into a stochastic policy suitable for policy gradient methods.

**GRPO.** Flow-GRPO [29] formulates the denoising process as a Markov Decision Process (MDP). Given a condition $\mathbf{c}$, a group of $G$ trajectories $\{\mathbf{x}^i\}_{i=1}^G$ is sampled. The advantage $\hat{A}_t^i$ for the $i$-th trajectory is estimated by normalizing the reward $R(\mathbf{x}_0^i, \mathbf{c})$ against the group statistics:

$$\hat{A}_t^i = \frac{R(\boldsymbol{x}_0^i, \boldsymbol{c}) - \mathrm{mean}(\{R(\boldsymbol{x}_0^i, \boldsymbol{c})\}_{i=1}^G)}{\mathrm{std}(\{R(\boldsymbol{x}_0^i, \boldsymbol{c})\}_{i=1}^G)}. \tag{3}$$

The model is then updated by maximizing the GRPO objective, which incorporates a KL-divergence constraint to prevent deviation from the reference policy $\pi_{\mathrm{ref}}$:

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathbf{c},\{\mathbf{x}^i\}}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathcal{L}_{\mathrm{clip}}(r_t^i, \hat{A}_t^i) - \beta D_{\mathrm{KL}}(\pi_\theta || \pi_{\mathrm{ref}})\right)\right], \tag{4}$$

where $\mathcal{L}_{\mathrm{clip}}$ denotes the standard PPO-style clipped surrogate objective, and $r_t^i$ is the probability ratio between the current and old policies. Furthermore, Flow-GRPO employs a denoise reduction strategy, using fewer timesteps during training to accelerate convergence without sacrificing inference quality.

## 4 Methodology

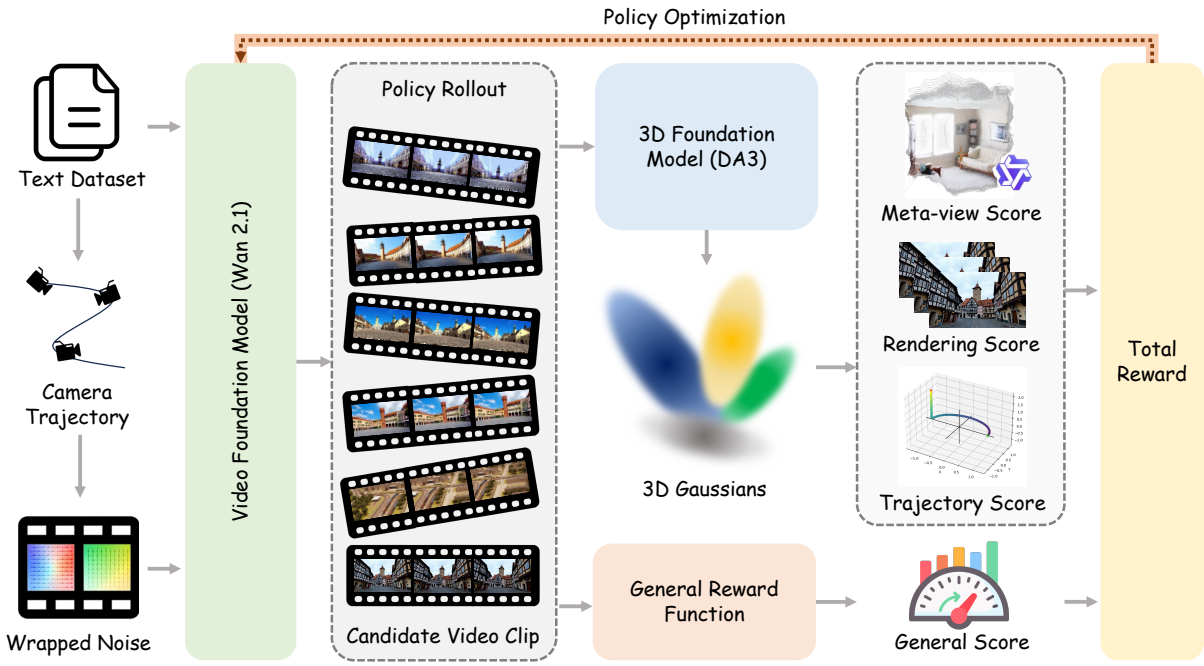


**Figure 2** **Overview of World-R1.** *Left:* Given a text prompt, we synthesize a camera trajectory and embed it into the latent space via noise wrapping [40] to achieve implicit camera conditioning. *Middle:* The foundation model [3] generates candidate video clips. *Right:* We employ an comprehensive reward system: the generated video is lifted to a 3D Gaussian Splatting representation via a 3D Foundation Model [20]. We compute a 3D-aware reward score based on meta-view evaluation, rendering fidelity, and trajectory alignment, combined with a general generation reward. The model is optimized using Flow-GRPO-Fast [29] to internalize these geometric constraints.

## 4.1 Overview

In this section, we detail the proposed framework for injecting world-modeling capabilities into a pre-trained video generation model [3]. The key distinction of our approach lies in its avoidance of explicit architectural modifications and its independence from specialized 3D-aware datasets, instead relying on noise manipulation and reinforcement learning.

## 4.2 Camera Conditioning

Unlike previous methods that require training auxiliary networks [32, 37] to encode camera poses, we adopt a parameter-free, implicit conditioning strategy inspired by Go-with-the-Flow [40]. Motivated by the observation that the initial noise distribution in diffusion models significantly influences the trajectory of the generated content, we embed camera motion priors directly into the latent initialization.

**Prompt-Driven Trajectory Generation.** We first map the prompt $c$ to a deterministic sequence of camera extrinsic matrices $E = \{E_t\}_{t=0}^{N}$, where $E_t$ represents the camera pose in 3D space at frame $t$. We define a keyword detection function $\phi(c)$ that scans the input prompt for predefined motion tokens $\mathcal{K} = \{\text{'push in'}, \text{'pan left'}, \text{'orbit left'}, \dots\}$. Based on the detected tokens, we instantiate a trajectory using a parameterized generator. Let $E_0 = I_{4\times4}$ be the canonical starting pose. The pose at step $t$ is computed recursively:

$$E_t = E_{t-1} \cdot T_{\text{action}}(t), \tag{5}$$

where $T_{\text{action}}$ denotes the transformation matrix specific to the motion type. When multiple camera movement instructions are present in the prompt, we concatenate the generated trajectories to obtain the final trajectory.

**Trajectory-to-Flow Projection.** To warp the noise representations, we project the 3D camera trajectory $E$ into a sequence of 2D dense optical flow fields. We adopt a pinhole camera model and approximate the scene geometry as a fronto-parallel plane located at a constant reference depth $z_{\text{ref}}$. For a pixel $u \in \mathbb{R}^2$ in frame $t$, its projected coordinate $u'$ in frame $t+1$ is derived via the planar homography induced by the relative camera motion:

$$u' \sim K \left( R_{\text{rel}} + \frac{1}{z_{\text{ref}}} \mathbf{t}_{\text{rel}} \mathbf{n}^\top \right) K^{-1} u, \tag{6}$$

where $K$ is the intrinsic matrix, $(R_{\text{rel}}, \mathbf{t}_{\text{rel}})$ denotes the relative rigid body transformation $E_{t+1}E_t^{-1}$, and $\mathbf{n} = [0, 0, 1]^\top$ is the normal vector of the image plane. This process yields the forward optical flow field defined as $f(u) = u' - u$.

**Discrete Noise Transport.** While the projected flow field $f(u)$ defines a continuous motion prior, the initial noise is defined on a discrete spatial grid. Directly warping noise using $f(u)$ would lead to variance collapse in overlapping regions and missing noise in disoccluded areas, violating the standard Normal distribution required by diffusion-based models. To address this, we adopt the discrete noise transport mechanism from Go-with-the-Flow [40], which formulates noise warping as a mass transport problem on a bipartite graph induced by the flow field.

Specifically, the continuous flow $f(u)$ induces discrete correspondences $v \to v'$ between source pixels $v$ at frame $t$ and target pixels $v'$ at frame $t+1$. Noise values are aggregated according to these correspondences, while a density tracker $\rho(v')$ records the number of incoming contributions. The transported noise is then normalized to preserve unit variance:

$$z_{t+1}(v') = \frac{1}{\sqrt{\rho(v')}} \sum_{v \to v'} z_t(v), \tag{7}$$

where $z_{t+1}$ is the noise value at a target pixel $v'$. This discrete transport scheme ensures that camera-induced spatial structure is injected into the initial noise while maintaining its standard Normal distribution across frames.

### 4.3 Reward Design

To ensure the generated videos enhance 3D capabilities while maintaining high visual fidelity, we design a composite reward mechanism. This objective function $R$ is formulated as a weighted aggregation of a physics-grounded 3D consistency term $R_{\text{3D}}$ and a general quality assessment term $R_{\text{gen}}$:

$$R(\mathbf{x}, c) = R_{\text{3D}}(\mathbf{x}, E, c) + \lambda_{\text{gen}} R_{\text{gen}}(\mathbf{x}, c). \tag{8}$$

Here $\mathbf{x}$ represents the generated video, $E$ denotes the target camera trajectory derived from the prompt condition $\mathbf{c}$, and $\lambda_{\text{gen}}$ is a balancing hyper-parameter.

**3D-Aware Reward.** We employ an analysis-by-synthesis strategy to distinguish between genuine 3D parallax and 2D content drift. We leverage the Depth Anything 3 [20] which directly reconstructs the scene geometry as a 3D Gaussian Splatting (3DGS) representation $\Phi_{\text{GS}}$ and estimates the corresponding camera trajectory $\hat{E}$ from the generated video $\mathbf{x}$. We define the 3D-aware reward $R_{\text{3D}}$ as a linear combination of three distinct consistency metrics tailored to geometry, appearance, and control accuracy:

$$R_{\text{3D}} = \mathcal{S}_{\text{meta}} + \mathcal{S}_{\text{recon}} + \mathcal{S}_{\text{traj}}. \tag{9}$$

Specifically, the geometric integrity term $\mathcal{S}_{\text{meta}}$ is computed by rendering 3D Gaussians from a novel meta-view. The renderings are then evaluated by Qwen3-VL [25] to assess text fidelity and structural reliability, effectively penalizing artifacts that remain occluded in the canonical view but indicate underlying geometric flaws. To ensure visual consistency, $\mathcal{S}_{\text{recon}}$ measures pixel-level fidelity by comparing the generated video $\mathbf{x}$ against its re-rendered counterpart from the 3DGS

representation $\Phi_{\text{GS}}$, quantifying the similarity via the negated perceptual distance ($1-$LPIPS [48]). Finally, the trajectory alignment term $\mathcal{S}_{\text{traj}}$ assesses control precision by calculating the deviation between the generated camera condition $E$ and the predicted trajectory $\hat{E}$, thereby ensuring the generated camera motion adheres strictly to the user instructions. For more details of the reward, please refer to Section A.1.

**General Generation Reward.** We further incorporate a general reward function $R_{\text{gen}}$ to guarantee the visual quality and aesthetic appeal of the generated content. We formulate this objective as the average aesthetic assessment across the first $K$ frames:

$$R_{\text{gen}}(\mathbf{x}) = \frac{1}{K}\sum_{t=0}^{K-1}\mathcal{H}(\mathbf{x}_t), \tag{10}$$

where $\mathcal{H}$ denotes the HPSv3 [49] score evaluated on each frame $\mathbf{x}_t$. By maximizing this aggregated score, we ensure the video consistently aligns with human visual preferences and maintains high perceptual standards throughout the sequence.

### 4.4 Dataset Preparation

A significant bottleneck in previous camera-control research has been the reliance on open-domain video datasets, where limited resolution and noisy text-video alignment often hinder precise control learning. To overcome this limitation and dissociate the learning of 3D constraints from specific video distributions, we construct a *Pure Text Dataset* specifically tailored for world simulation.

We leverage the advanced instruction-following and creative capabilities of Gemini [50, 51] to synthesize a diverse corpus of high-quality scene descriptions. The core advantage of this dataset lies in its independence from fixed visual priors, allowing the model to synthesize a vast array of distinct scenarios, ranging from natural landscapes and urban structures to surrealist environments. The dataset comprises approximately 3,000 unique entries, systematically categorized to cover a wide spectrum of visual domains and physical properties. Furthermore, we categorize these prompts by control complexity, including samples with implicit motion, single directional commands, and complex composite trajectories. This multi-level structure enables the model to learn physics-compliant generation across varying degrees of difficulty. For more details about the dataset, please refer to Section B.

### 4.5 Training Strategy

Strict adherence to 3D consistency can inadvertently suppress the generation of non-rigid dynamics, such as moving characters or naturally deforming objects. To balance geometric fidelity with the capability for dynamic scene generation, we adopt a periodic decoupled training strategy.

Our training datasets contain approximately 500 prompts specifically describing highly dynamic scenes. During training, we implement a multi-stage cycle. In the primary stage, the model is trained with the full weighted reward to enforce 3D-aware capability. However, to prevent the model from overfitting to static geometric constraints, we introduce a dynamic fine-tuning phase after every 100 training steps. In this phase, we temporarily disable the 3D-aware reward $R_{\text{3D}}$ and optimize the model exclusively on the dynamic data subset using only the general reward $R_{\text{gen}}$. This mechanism acts as a regularizer, ensuring the model retains its generalization power for complex dynamic motions while learning world simulation. The efficacy of this strategy is verified in Section 5.3. Visualization results are illustrated in Section E.

## 5 Experiments

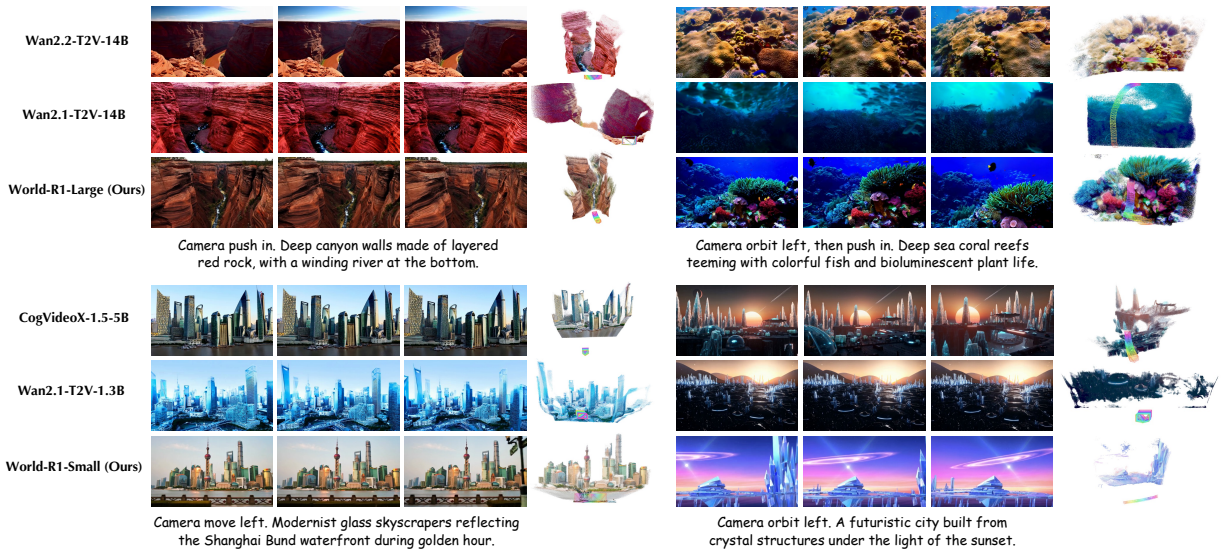


**Figure 3** **Qualitative Comparison.** We visualize the generated video frames and the corresponding 3D reconstruction for diverse scenes involving different camera motions. Compared to the video foundation models [1, 3], World-R1 demonstrates superior 3D consistency. Our model preserves strict scene permanence and geometry performance, yielding dense and structured 3D visualizations that confirm its robust world-modeling capabilities.

## 5.1 Experimental Settings

**Implementation Details.** We leverage the state-of-the-art open-source video generation models Wan 2.1 1.3B and Wan 2.1 14B [3] as backbone foundations. By fine-tuning on our proprietary dataset (Section 4.4), we obtain two model variants. *World-R1-Small* is trained using 48 NVIDIA H200 GPUs, while *World-R1-Large* is trained using 96 NVIDIA H200 GPUs. Both models are optimized at a video resolution of $832 \times 480$. For reinforcement learning, we adopt Flow-GRPO-Fast [29] across both experimental settings, utilizing 48 parallel groups with a group size of 8 to ensure training stability and scalability. For more details, please refer to Section A.2.

**Evaluation Metrics.** We employ a dual-pronged evaluation protocol focusing on both geometric integrity and general generative quality to assess world-modeling capabilities. We first introduce a reconstruction-based metric rooted in 3D geometry to quantify 3D consistency. We employ 3DGS [52, 53] to reconstruct the underlying scene geometry for each generated video. We then re-render the scene from the exact camera trajectory specified in the input prompt. We measure the extent to which the video maintains structural coherence by calculating the PSNR, SSIM [54], and LPIPS [48] between the generated video and its 3D-consistent re-rendering. High reconstruction fidelity indicates that the model has successfully suppressed geometric hallucinations and temporal warping. To reduce dependence on the reconstruction pipeline, we additionally report a reconstruction-independent Multi-View Consistency Score (MVCS) in Section D. Furthermore, we utilized VBench [55] as a standardized benchmark to evaluate general video quality. We report a comprehensive suite of sub-metrics including Aesthetic Quality, Imaging Quality, Motion Smoothness, and Consistency scores. This ensures that our alignment process improves the visual fidelity and dynamic range of the base foundation model.

## 5.2 SoTA Comparisons and Analysis

We benchmark World-R1 against leading foundational video generation models including Wan 2.1 [3] and CogVideoX [1]. We also compare our approach with state-of-the-art auxiliary control methods such as CameraCtrl [32] and ReCamMaster [37] to contextualize our contributions further. These methods introduce explicit architectural modules for camera guidance.

**Table 1** Evaluation of general video generation quality on VBench [55]. Due to resource constraints, we did not test the performance of World-R1-Large. Best scores are in blue, second-best in green, third-best in yellow.

| Method | Aesthetic Quality ↑ | Imaging Quality ↑ | Motion Smoothness ↑ | Subject Consistency ↑ | Background Consistency ↑ |
|---|---|---|---|---|---|
| CogVideoX-1.5-5B [1] | 62.07 | 65.34 | 98.15 | 96.56 | 96.81 |
| Wan2.1-T2V-1.3B [3] | 62.43 | 66.51 | 97.44 | 96.34 | 97.29 |
| GCD [56] | 38.21 | 41.56 | 98.37 | 88.94 | 92.00 |
| Trajectory-Attention [57] | 38.50 | 51.00 | 98.21 | 90.60 | 92.83 |
| DAS [58] | 39.86 | 51.55 | 99.14 | 90.34 | 92.03 |
| ReCamMaster [37] | 42.70 | 53.97 | 99.28 | 92.05 | 93.83 |
| **World-R1-Small (Ours)** | 65.74 | 67.53 | 98.55 | 97.58 | 96.67 |

**Table 2** **Quantitative comparison of 3D consistency.** We evaluate the structural stability of generated videos by reconstructing the scene geometry via 3DGS [52] and re-rendering it on the test set proposed in Section 4.4. Notably, World-R1 significantly outperforms baseline foundation models, achieving a substantial reduction in geometric hallucinations and improved 3D consistency.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| CogVideoX-1.5-5B [1] | 24.44 | 0.783 | 0.242 |
| Wan2.2-T2V-14B [3] | 23.47 | 0.779 | 0.253 |
| Wan2.2-T2V-5B [3] | 22.36 | 0.716 | 0.303 |
| Wan2.1-T2V-14B [3] | 19.76 | 0.629 | 0.405 |
| Wan2.1-T2V-1.3B [3] | 17.40 | 0.550 | 0.467 |
| **World-R1-Small (Ours)** | 27.63 | 0.858 | 0.201 |
| **World-R1-Large (Ours)** | **27.67** | **0.865** | **0.162** |

**Quantitative Analysis.** Table 2 presents the performance of World-R1 on 3D consistency. Our approach significantly outperforms all baselines and achieves a PSNR improvement of **10.23dB** and **7.91dB** along with a substantial improvement in SSIM and LPIPS. This quantitative leap confirms that our RL-based alignment successfully injects rigid geometric constraints into the video generation process and mitigates the artifacts often observed in standard foundation models. Auxiliary control methods such as CameraCtrl [32] often suffer from texture distortion and lead to lower reconstruction scores despite adhering to trajectories.

Table 1 details the performance on VBench [55]. World-R1 surpasses the performance of the original Wan 2.1 [3] backbone. Meanwhile, compared to models controlled by explicit cameras, we far surpass them in Aesthetic Quality, Imaging Quality, and Subject Consistency. Section D provides additional robustness analyses, including dataset scaling, 121-frame evaluation, category-wise results for complex scenes, and a consolidated comparison with 3D-conditioned and camera-control methods. These results demonstrate that our framework effectively prevents the degradation of visual quality often associated with heavy structural constraints.

**Qualitative Analysis.** Visual comparisons presented in Figure 3 illustrate the capability gap. Baseline models frequently exhibit unrealistic transitions where objects vanish or walls warp in scenarios requiring complex camera motions such as circling a building or pushing through a corridor. World-R1 maintains strict object permanence and rigid geometry in contrast. The 3DGS reconstruction visualizations further substantiate this finding. The point clouds derived from our videos are dense and structured, while reconstructions from baseline videos appear sparse and noisy due to multi-view inconsistencies.

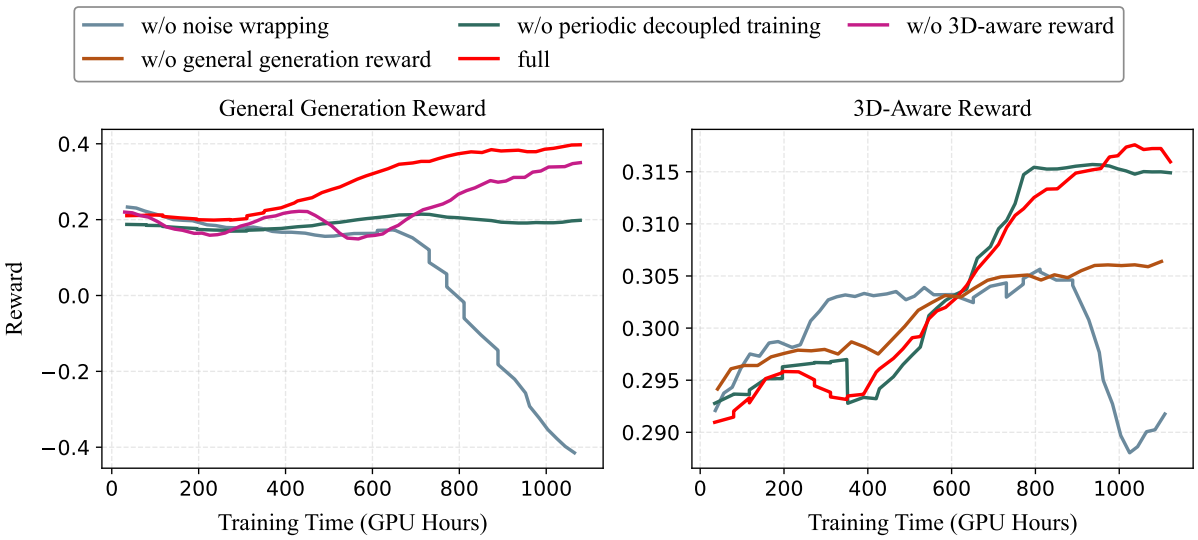


**Figure 4** **Ablation study.** We visualize the evolution of the general generation reward and the 3D-aware reward during fine-tuning of World-R1-Small. The curves demonstrate the contribution of each component.

**User Study.** To evaluate human subjective preference, we conducted a blind user study comparing World-R1 against the base model Wan 2.1 [3] (specifically, Wan2.1-T2V-1.3B vs. World-R1-Small and Wan2.1-T2V-14B vs. World-R1-Large). We curated a test set of 30 complex prompts covering various domains (e.g., architectural fly-throughs, natural landscapes) and camera motions (e.g., push in, orbit). 25 participants were presented with randomized side-by-side video pairs and asked to select the superior model based on three criteria: (1) Geometric Consistency (stability of objects and structures), (2) Camera Control Accuracy (adherence to the specified trajectory), and (3) Overall Visual Quality. World-R1 demonstrates a commanding lead in both geometric consistency and camera control accuracy, achieving win rates of 92% and 76% respectively over Wan 2.1 [3]. Notably, in terms of overall preference, World-R1 was chosen as the best video in 86% of comparisons, validating that our RL-based alignment effectively bridges the gap between creative generation and geometric reasonableness.

## 5.3 Ablation Study

We perform ablation studies on World-R1-Small to verify the effectiveness of our key components as shown in Figure 4, systematically isolating the contributions of our reward functions, model conditioning, and data training strategy. For **reward mechanism**, our analysis confirms that the 3D-aware reward $R_{3D}$ is fundamental for establishing geometric consistency, whereas the general generation reward $R_{gen}$ proves indispensable for maintaining high perceptual fidelity and preventing the aesthetic degradation. For **model conditioning**, we observed that removing the implicit camera injection via noise wrapping leads to significantly slower convergence and inferior trajectory alignment, demonstrating that embedding motion priors directly into the latent initialization provides a critical inductive bias for the optimization process. For **training strategy**, without the intermittent relaxation of constraints using the dynamic data subset, the model overfits to static rigidity and suppresses natural non-rigid dynamics, thereby validating our approach of using dynamic fine-tuning as a regularizer to balance strict world simulation with fluid motion generation. Section D.7 further analyzes reward hacking and reports component-wise reward ablations.

## 6 Conclusion

We presented World-R1, a scalable paradigm for endowing video generation models with robust world-modeling capabilities. By reformulating the alignment of video generation and 3D geometry as a reinforcement learning problem, we successfully elicit latent spatial awareness from pre-trained models without requiring explicit 3D architectural modules or expensive supervised datasets. Our approach leverages a composite reward system grounded in multi-view consistency and semantic coherence, optimized via Flow-GRPO, to ensure physical validity while maintaining visual fidelity. The proposed implicit camera conditioning and periodic training strategy further enable precise trajectory control without compromising the generation of dynamic non-rigid content. Quantitative and qualitative evaluations confirm that our method significantly improves geometric consistency and preserves general video quality compared to state-of-the-art baselines. By effectively transforming video generators into geometrically consistent simulators, our framework opens new avenues for applications in autonomous driving simulation and physical world modeling.

**Limitations and Future Work.** While World-R1 successfully injects world-modeling capabilities into video foundation models, two limitations remain. First, the computational cost of applying reinforcement learning to video generation is still a significant bottleneck. Unlike supervised fine-tuning, online RL requires repeated video rollouts and reward evaluation, making the training process more expensive than standard post-training pipelines. Developing more efficient rollout strategies, lower-cost reward evaluation, and stable video RL optimization will therefore be important directions for future work. Second, World-R1 is built on top of existing video foundation models and is consequently bounded by their generative capacity. Challenging cases such as dense multi-object composition, fine-grained non-rigid motion, detailed hand dynamics, and very long-horizon scene evolution may still inherit artifacts from the base model. As stronger video foundation models become available, our post-training framework can directly benefit from their improved scene understanding and motion generation capabilities.

## Impact Statement

This paper introduces a method to improve the geometric consistency of video generation models by aligning video synthesis with 3D constraints through reinforcement learning and flow-matching optimization. This advancement enhances the reliability of generated videos for applications requiring high physical accuracy, such as physical world modeling and autonomous driving simulation. The societal and ethical considerations of this work align with those of existing high-fidelity video generation techniques. The approach does not introduce novel model architectures or data modalities that would create new ethical risks beyond known issues like the potential misuse of synthetic content. By utilizing a fast training strategy and a text-based dataset, the method offers a computationally efficient alternative, potentially reducing the carbon footprint associated with large-scale model training. No additional broader impacts requiring specific discussion are identified.

## Author Contributions

Weijie Wang led the overall project, including the end-to-end framework design, code implementation, paper writing, and experimental validation. Xiaoxuan He was responsible for large-scale experiment debugging, validation, and experimental optimization. Youping Gu contributed to the development and debugging of the camera conditioning module and conducted part of the testing experiments. Zeyu Zhang, Yefei He, Yanbo Ding and Xirui Hu contributed to experimental analysis, evaluation design and manuscript refinement. Yifan Yang, Donny Y. Chen, Zhiyuan He, Yuqing Yang and Bohan Zhuang advised the project, provided high-level research guidance, supervised the study, and contributed to the final manuscript revision.

## References


[1] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *Int. Conf. Learn. Represent.*, 2025.

[2] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7310–7320, 2024.

[6] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7038–7048, 2024.

[7] Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In *Int. Conf. Comput. Vis.*, pages 27326–27337, 2025.

[8] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025.

[9] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

[10] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

[11] Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. *arXiv preprint arXiv:2509.15130*, 2025.

[12] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Int. Conf. Comput. Vis.*, 2025.

[13] Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.

[14] Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6078–6088, 2025.

[15] Weijie Wang, Jiagang Zhu, Zeyu Zhang, Xiaofeng Wang, Zheng Zhu, Guosheng Zhao, Chaojun Ni, Haoxiao Wang, Guan Huang, Xinze Chen, et al. Drivegen3d: Boosting feed-forward driving scene generation with efficient video diffusion. *arXiv preprint arXiv:2510.15264*, 2025.

[16] Zixuan Huang, Xiang Li, Zhaoyang Lv, and James M. Rehg. How much 3d do video foundation models encode? *arXiv preprint arXiv:2512.19949*, 2025.

[17] Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward. *arXiv preprint arXiv:2603.26599*, 2026.

[18] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[19] Weijie Wang, Qihang Cao, Sensen Gao, Donny Y Chen, Haofei Xu, Wenjing Bian, Songyou Peng, Tat-Jen Cham, Chuanxia Zheng, Andreas Geiger, et al. Feed-forward 3d scene modeling: A problem-driven perspective. *arXiv preprint arXiv:2604.14025*, 2026.

[20] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[21] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Adv. Neural Inform. Process. Syst.*, volume 37, pages 21875–21911, 2024.

[22] Weijie Wang, Donny Y Chen, Zeyu Zhang, Duochao Shi, Akide Liu, and Bohan Zhuang. Zpressor: Bottleneck-aware compression for scalable feed-forward 3dgs. *Advances in Neural Information Processing Systems*, 38:113407–113436, 2026.

[23] Weijie Wang, Yeqing Chen, Zeyu Zhang, Hengyu Liu, Haoxiao Wang, Zhiyuan Feng, Wenkang Qin, Feng Chen, Zheng Zhu, Donny Y Chen, et al. Volsplat: Rethinking feed-forward 3d gaussian splatting with voxel-aligned prediction. *arXiv preprint arXiv:2509.19297*, 2025.

[24] Duochao Shi, Weijie Wang, Donny Y Chen, Zeyu Zhang, Jia-Wang Bian, Bohan Zhuang, and Chunhua Shen. Revisiting depth representations for feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2506.05327*, 2025.

[25] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

[27] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[28] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[29] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. In *Adv. Neural Inform. Process. Syst.*, 2025.

[30] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[31] 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 Conf. Comput. Graph. Interact. Tech.*, pages 1–11, 2024.

[32] 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.

[33] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Adv. Neural Inf. Process. Syst.*, 37:34322–34348, 2024.

[34] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. *Adv. Neural Inform. Process. Syst.*, 37:16240–16271, 2024.

[35] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025.

[36] Sixiao Zheng, Zimian Peng, Yanpeng Zhou, Yi Zhu, Hang Xu, Xiangru Huang, and Yanwei Fu. Vidcraft3: Camera, object, and lighting control for image-to-video generation. *arXiv preprint arXiv:2502.07531*, 2025.

[37] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Int. Conf. Comput. Vis.*, 2025.

[38] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Int. Conf. Comput. Vis.*, pages 28785–28796, 2025.

[39] Cheng Zhang, Hanwen Liang, Donny Y Chen, Qianyi Wu, Konstantinos N Plataniotis, Camilo Cruz Gambardella, and Jianfei Cai. Panflow: Decoupled motion control for panoramic video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12385–12393, 2026.

[40] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13–23, 2025.

[41] Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. *arXiv preprint arXiv:2512.15716*, 2025.

[42] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21685–21695, 2025.

[43] Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

[44] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4): 1–12, 2018.

[45] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 22160–22169, 2024.

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[47] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 586–595, 2018.

[49] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Int. Conf. Comput. Vis.*, pages 15086–15095, 2025.

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[51] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[52] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[53] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et al. gsplat: An open-source library for gaussian splatting. *JMLR*, 26(34):1–17, 2025.

[54] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[55] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21807–21818, 2024.

[56] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *Eur. Conf. Comput. Vis.*, pages 313–331. Springer, 2024.

[57] Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. *arXiv preprint arXiv:2411.19324*, 2024.

[58] 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.

[59] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Int. Conf. Comput. Vis.*, 2025.

[60] 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.

[61] Yawen Luo, Jianhong Bai, Xiaoyu Shi, Menghan Xia, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Tianfan Xue. Camclonemaster: Enabling reference-based camera control for video generation. *arXiv preprint arXiv:2506.03140*, 2025.

[62] Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *arXiv preprint arXiv:2512.03453*, 2025.

[63] 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.

[64] Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. Flashworld: High-quality 3d scene generation within seconds. *arXiv preprint arXiv:2510.13678*, 2025.

[65] Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. *arXiv preprint arXiv:2601.05138*, 2026.

# A Implementation Details

## A.1 Reward Formulation and Details

The core of our alignment strategy is the 3D-aware reward, which utilizes an analysis-by-synthesis approach to penalize geometric inconsistencies. Given a generated video clip $x$, we employ Depth Anything 3 to lift the sequence into a 3D Gaussian Splatting (3DGS) representation, denoted as $\Phi_{GS}$, and simultaneously estimate the camera trajectory $\hat{E}$. The reward $R_{3D}$ is computed as the linear combination of three distinct metrics as in Equation (9): $R_{3D} = \mathcal{S}_{meta} + \mathcal{S}_{recon} + \mathcal{S}_{traj}$, all rewards are limited to $[0, 1]$ and added directly. For the final score calculation, the value range of $R_{gen}$ is $[-1, 1]$, and the value range of $R_{3D}$ is $[0, 3]$. We will add them directly with $\lambda_{\text{gen}} = 1$ according to Equation (8).

The **Geometric Integrity Score ($\mathcal{S}_{meta}$)** is the most critical component for detecting hallucinations that appear plausible in 2D but fail in 3D (e.g., flat "billboard" objects). We render the optimized $\Phi_{GS}$ from a novel *meta-view*. a camera pose significantly offset from the generation trajectory (e.g., backward from the origin). From this perspective, we can have a more comprehensive observation of the whole scene. We employ Qwen3-VL [25] as a semantic critic to evaluate the structural plausibility of this meta-view. The VLM is prompted to act as a graphics expert, identifying artifacts such as floaters, geometric distortion, or texture stretching. The specific system instruction provided to the VLM is detailed below:

**System Instruction:** You are a professional 3D vision expert. I used a text prompt to generate a video and reconstructed a corresponding 3D Pointmap from the video.
**Original Prompt:** [Text Prompt]
**Criteria:** Your task is to judge the quality of the original video by analyzing the provided image of its resulting 3D pointmap. A good video (smooth, orbiting camera) creates a good pointmap. A bad video (static, jittery, or zooming) creates a bad pointmap. Please provide a score from 0 to 9 based on these criteria:

- 9: Excellent - A dense, clean, and complete 3D model. Perfect 360° orbital motion, high stability.
- 7-8: Good - A clear object with strong 3D structure. May have minor holes or noise. Good, smooth camera arc with strong parallax.
- 4-6: Mediocre - Object is recognizable, but the map is sparse, noisy, or "flat" (lacks 3D depth). Poor parallax (e.g., just a zoom or pan instead of an orbit), or the video was jittery, blurry, or had object/lighting inconsistencies.
- 2-3: Poor - A chaotic jumble of points or a simple 2D projection. Static camera (no motion) or completely unstable.
- 0-1: Very Poor - Empty or just random noise. Unusable.

**Scoring:** Output only a single digit (0-9):

To ensure the meta-view evaluation is robust, we visualize the difference between the good meta-view and bad meta-view in Figure A. The canonical view usually appears structurally sound, whereas the meta-view reveals the "cardboard prop" effect in baseline models, which our reward explicitly suppresses. We multiply the obtained score by 0.1 to obtain the final score, ensuring the score falls within the range $[0, 1]$.

The **Reconstruction Fidelity Score ($\mathcal{S}_{recon}$)** ensures that the lifted 3D representation remains faithful to the original generated content. We re-render the scene from the estimated trajectory $\hat{E}$ to obtain the reconstruction $\hat{x}$ and measure the pixel-level similarity using LPIPS. This term is calculated as $\mathcal{S}_{recon} = 1 - \text{LPIPS}(x, \hat{x})$, favoring generations that allow for photorealistic 3D reconstruction.

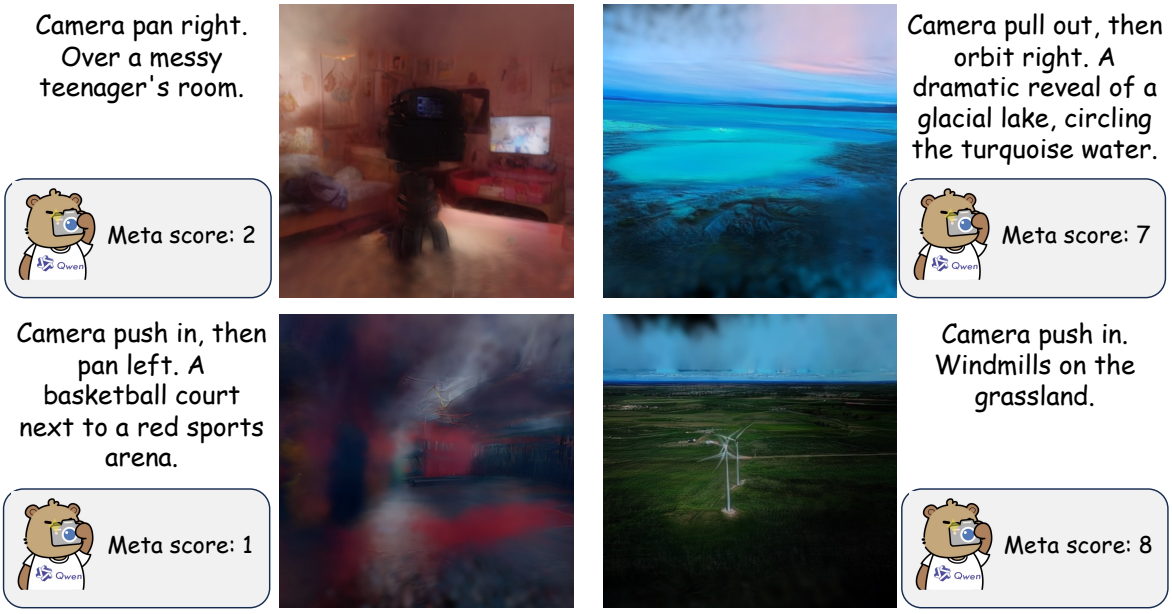


**Figure A** **Meta view visualization.** Left: Due to unstable camera trajectory and poor 3D consistency, the reconstructed 3DGS cannot be displayed stably. On the right: A high-quality video with a good meta view and consistent with the prompt description.

The **Trajectory Alignment Score** ($\mathcal{S}_{traj}$) quantifies the controllability of the generation. It measures the deviation between the user-specified input camera trajectory $E$ (embedded via our implicit conditioning) and the actual camera motion $\hat{E}$ derived from the video. We compute this as the negative exponential of the $L_2$ distance for translation and the geodesic distance for rotation, ensuring the generated camera movement strictly adheres to the prompt instructions.

## A.2 More Training Details

**Infrastructure and Base Models.** Our experimental framework is built upon the state-of-the-art Wan 2.1 video foundation models [3]. We train two variants of our model: *World-R1-Small*, initialized from Wan2.1-T2V-1.3B, and *World-R1-Large*, initialized from Wan2.1-T2V-14B. The training infrastructure leverages a high-performance compute cluster equipped with NVIDIA H200 GPUs. Specifically, the optimization of the Small variant is distributed across 48 H200 GPUs, while the Large variant requires a cluster of 96 H200 GPUs to handle the increased memory and computational demands. All training operations are performed at a video resolution of $832 \times 480$ to balance visual fidelity with computational throughput during the reinforcement learning rollout phase.

**Optimization and Rollout.** We employ the Flow-GRPO-Fast algorithm [29] to align the video generation policy. Unlike standard supervised fine-tuning, this online reinforcement learning approach requires generating full video sequences(rollouts) during training to compute the reward signals. For these rollouts, we utilize a denoising schedule to ensure the generated samples are of sufficient quality for the 3D Gaussian Splatting reconstruction process. To stabilize the policy gradient estimation and reduce variance, we configure the GRPO with a group size of $G = 8$. The training is scaled with a total batch size distributed across 48 parallel groups, ensuring diverse trajectory sampling for accurate advantage estimation.

**Periodic Decoupled Training Strategy.** A critical challenge in reinforcing geometric constraints is the potential suppression of non-rigid dynamics, such as fluid motion or biological deformation. To mitigate this, we implement a Periodic Decoupled Training strategy that alternates between

geometric alignment and dynamic preservation. The training process is structured into cycles. In the primary phase, the model is optimized on the full mixed dataset using the complete composite reward function ($R_{3D} + \lambda_{gen} R_{gen}$) to enforce 3D consistency. However, to prevent the model from overfitting to static rigidity, we trigger a *dynamic fine-tuning phase* every 100 training steps. During this interval, we temporarily suspend the 3D-aware reward component $R_{3D}$ and optimize the policy exclusively on the *Dynamic Data Subset*–a curated collection of approximately 500 prompts describing high-entropy scenes (e.g., fire, flowing water, moving crowds). In this phase, the model is guided solely by the general generation reward $R_{gen}$. This periodic relaxation of geometric constraints acts as a regularization mechanism, ensuring that World-R1 retains the generalization power required to synthesize complex, dynamic environments while internalizing the laws of physics.

# B Dataset Details

To facilitate the training of robust world simulation capabilities, we constructed a proprietary Pure Text Dataset specifically tailored to decouple geometric constraint learning from the biases inherent in existing open-domain video distributions. This dataset, comprising approximately 3,000 unique entries, was synthesized using the advanced instruction-following and creative reasoning capabilities of Gemini [50, 51]. The data generation process was governed by a hierarchical prompt engineering strategy, ensuring a diverse coverage of visual domains and control complexities. Below, we detail the taxonomy of the camera control primitives and the semantic categories of the dataset, providing the complete list of prompt templates used for training.

## B.1 Generation Pipeline

The construction of our dataset was automated through a structured generation pipeline leveraging Gemini-3. To ensure the generated text-video pairs effectively disentangle geometric learning from visual bias, we designed a rigorous prompt engineering strategy. This pipeline consists of two core components: a role-playing system instruction for the LLM and a predefined action space for camera control.

**Gemini system instruction.** We instructed the LLM to function as an expert cinematographer, requiring it to synthesize scene descriptions that are not only visually rich but also physically grounded. The system prompt explicitly enforces the pairing of scene geometry with appropriate camera movements (e.g., assigning a `push_in` trajectory to a "tunnel" or "corridor" scene to maximize parallax effect). The exact system instruction used in our generation pipeline is presented below:

**Prompt for Dataset Generation**
**Role:** You are an expert Cinematographer and 3D Set Designer. **Objective:** Generate a dataset of high-fidelity text-to-video prompts that describe geometrically consistent 3D worlds. **Constraints:**

- **Physical Plausibility:** The scene must follow real-world physics (unless specified as Surrealism). Objects must have defined spatial relationships.
- **Camera-Scene Matching:** The assigned camera movement must logically fit the scene layout (e.g., use 'orbit' for centered objects, 'push_in' for depth exploration).
- **Diversity:** Cover the following domains: Natural Landscapes, Urban & Architecture, Micro World, and Fantasy.

**Action Space (Camera Movements):** Choose strictly from the following list: ['push_in', 'pull_out', 'move_left', 'move_right', 'orbit_left', 'orbit_right', 'pan_left', 'pan_right', 'pull_left', 'pull_right', 'fixed']
**Output Format (JSON):** { "prompt": "Full descriptive text...", "camera_logic": "Selected movement", "domain": "Category", "layout_type": "Intra/Inter/Static" }

**Camera trajectory definitions.** To standardize the action space, we use a comprehensive taxonomy of camera trajectories following WorldScore [59]. These are categorized by their topological effect on the scene navigation: *Intra-scene* (exploring within a volume), *Inter-scene* (transitioning across space), and *Static* (stationary observation). The precise definition for each trajectory primitive is detailed below:

**Camera Trajectory Taxonomy**

**1. Intra-scene Exploration**

*Movements that investigate the depth and 3D structure of a specific subject.*

- **push_in**: Move forward along the optical axis. (Focus: Depth & Parallax)
- **orbit_left**: Revolve counter-clockwise around a focal point. (Focus: 360° Object Consistency)
- **orbit_right**: Revolve clockwise around a focal point. (Focus: 360° Object Consistency)

**2. Inter-scene Transition**

*Movements that shift the viewport to reveal new environments or expand context.*

- **pull_out**: Move backward along the optical axis. (Focus: Context Reveal)
- **move_left**: Lateral truck left. (Focus: Parallax)
- **move_right**: Lateral truck right. (Focus: Parallax)
- **pan_left**: Rotate camera yaw left on axis. (Focus: Panoramic View)
- **pan_right**: Rotate camera yaw right on axis. (Focus: Panoramic View)

**3. Composite Trajectories**

*Complex multi-axis maneuvers testing long-horizon consistency.*

- **pull_left**: Sequence: `move_left` → `pull_out` → `pan_left`.
- **pull_right**: Sequence: `move_right` → `pull_out` → `pan_right`.

**4. Static Observation**

*Stationary camera to isolate temporal dynamics.*

- **fixed**: No ego-motion. (Focus: Fluid/Particle Dynamics)

## B.2 Dataset Taxonomy and Examples

### B.2.1 Natural Landscapes

This category focuses on large-scale rigid geometry and natural fluid dynamics. It tests the model's ability to maintain consistency across vast distances and handle organic structures. The prompts are divided into Landforms, Water Features, and Weather & Time.

**Landforms.** Prompts in this sub-category focus on geological structures and complex terrain traversal.

**Prompts:**
A rugged mountain pass covered in scree slopes and hardy alpine flowers.
Giant basalt columns forming a hexagonal cliff face by the sea.
A lush, hidden valley completely enclosed by towering cliffs.
A geothermal area with bubbling mud pools and steaming vents in a rocky landscape.
Camera orbit left, transitioning into a pan right. Circling a towering hoodoo, then sweeping across the canyon landscape.
Camera pan right, then pull out. A wide view of terraced rice paddies, pulling back to show the entire mountain slope.
Camera move left, pull out, then pan right. Starting in a deep gorge, revealing the scale of the cliffs above.

**Water features.** These prompts challenge the model with liquid physics, reflections, and transparency.

**Prompts:**
A turquoise lagoon surrounded by white sand beaches and palm trees.
Massive ocean waves crashing onto a rocky shore during a storm.
Camera move right. Past a frozen waterfall with giant icicles.
Camera orbit left. An aerial view circling a series of interlocking lakes.
Camera orbit right. Around a large rock formation in the sea.
Camera pan left. Scanning the surface of a misty lake.
Camera push in, then pan right. Moving towards a misty lake, then looking across the water.
Camera pull out, then orbit right. A dramatic reveal of a coastal archway, circling the rock formation.
Camera move right, then push in. Along a bubbling brook, then focusing on a small fish.

**Weather & time.** These entries evaluate atmospheric rendering and dynamic lighting conditions.

**Prompts:**
A dramatic cloudscape at sunset, with towering cumulus clouds catching the light.
A peaceful, starry night over a calm mountain lake, with stars reflected in the water.
A foggy coastline with the sound of a foghorn in the distance.
Camera fixed. A dramatic view of a sunset reflection on wet sand.
Camera push in. Into the pure blackness of a moonless night in the wilderness.
Camera pull out. Revealing a hazy jungle canopy at sunrise.
Camera move left. Watch lightning bolts strike during a night storm.
Camera move right. Across a desaturated landscape on an overcast day.
Camera orbit left. Around a sunrise with sun pillars.
Camera orbit right. Circling trees with leaves blowing on a windy autumn day.
Camera pan left. Across a twilight desert sky with color gradients.
Camera pan right. Following falling snow in a forest at night.
Camera fixed. A view of crepuscular rays over the ocean.
Camera push in. Towards a crisp, clean landscape after a rainstorm.
Camera pull out. Revealing low-lying fog blanketing a field at dawn.
Camera pan left, then push in. Scanning a hazy jungle sunrise before focusing on the sun.
Camera move left, pull out, then pan left. Starting with lightning striking, revealing the thunderstorm at night.
Camera push in, then pan right. Moving towards an overcast landscape, then scanning the flat light.
Camera pull out, then orbit right. A dramatic reveal of a sunrise with sun pillars, circling the light phenomenon.
Camera move right, then push in. Along windy autumn trees, focusing on a blowing leaf.

### B.2.2 Urban and Architectural

The Urban and Architectural domain emphasizes strict perspective correctness, vanishing points, and the preservation of straight lines.
**Urban landscapes.**

**Prompts:**
A futuristic cyberpunk city street drenched in neon rain at night.
A charming cobblestone street in a medieval European town at sunrise.
The skyline of a modern metropolis reflecting in a calm bay at dusk.
The historic Bund in Shanghai illuminated with golden lights at night.
A post-apocalyptic city ruin reclaimed by nature, with vines growing on skyscrapers.
A busy intersection in Tokyo filled with pedestrians and colorful signs.
Camera pull out, then orbit left. A dramatic reveal of a floating city, circling the platform.
Camera move right, then push in. Along a rainy street, then focusing on a reflection.
Camera orbit left, transitioning into a pan right. Circling a town square, then sweeping the crowd.
Camera pan right, then pull out. A wide view of traffic, pulling back to show the gridlock.
Camera move left, pull out, then pan right. Starting in a garden, revealing the suburban neighborhood.

**Indoor spaces.**

**Prompts:**
Camera pull out. Revealing a vast warehouse interior.
Camera move left. Along the lockers in a school hallway.
Camera move right. Past gaming machines in a vintage arcade.
Camera orbit left. Circling a plant in a greenhouse.
Camera orbit right. Around a table in a conference room.
Camera pan left. Across a luxury hotel lobby.
Camera pan right. Over the water in an indoor pool.
Camera fixed. A view of a rustic kitchen.
Camera push in. Into a dark corridor in an abandoned hospital.
Camera pull out. Revealing a packed casino floor.
Camera move left. Along the racks in a walk-in closet.

**Infrastructure.**

**Prompts:**
An empty, futuristic metro station with polished floors and blue lighting.
A lonely gas station on a desert highway illuminated by fluorescent lights.
A massive suspension bridge spanning a foggy strait.
Camera move left, then push in. Along a train platform, then approaching the track.
Camera orbit right, transitioning into a pull out. Circling a gas pump, then pulling back to show the station.
Camera pan left, then push in. Scanning a factory floor, then focusing on a machine.
Camera move right, pull out, then pan right. Starting at a suitcase, revealing the airport terminal.

### B.2.3 Micro and Still Life

Designed to evaluate depth-of-field handling and texture fidelity, this category focuses on small-scale objects and macro observation.

**Desktop still life.**

**Prompts:**
Camera pull out. Revealing a messy artist's desk.
Camera move left. Along the keys of a mechanical keyboard.
Camera move right. Past a row of antique books.
Camera orbit left. Circling a vase of flowers.
Camera orbit right. Around a burning candle flame.
Camera pan left. Scanning a table of food.
Camera pan right. Across a blueprint on a drafting table.
Camera fixed. A view of an hourglass with falling sand.
Camera push in. Towards the nib of a fountain pen.
Camera pull out. From a macro of a coin to a pile of money.
Camera move left. Along the spine of a leather notebook.
Camera move right. Past spools of colorful thread.
Camera push in, then orbit left. Approaching a feather quill, then circling the inkwell.
Camera pull out, then pan left. Revealing a record player from the needle.
Camera move right, then push in. Along a coaster, then focusing on a water ring.
Camera orbit right, then pull out. Circling a chocolate, then showing the box.
Camera pan left, then push in. Scanning a newspaper, then focusing on a headline.
Camera move left, pull out, then pan left. Starting at a screw, revealing the glasses.
Camera push in, then pan right. Moving towards a matcha whisk, then looking at the foam.

**Micro world.**

**Prompts:**
A close-up of a butterfly wing, revealing the mosaic of tiny scales.
The fuzzy texture of mold growing on a piece of bread.
A microscopic view of blood cells flowing through a capillary.
The jagged landscape of a vinyl record groove viewed under a microscope.
A grain of sand that looks like a jewel stone when magnified.
The delicate stamens of a hibiscus flower covered in yellow pollen.
A close-up of bubbles rising in a glass of carbonated soda.
The rough, cratered surface of a match stick head before ignition.
A spider web covered in tiny water droplets, glistening in the sun.
The fibers of a piece of paper, looking like a tangled forest.
A macro view of human iris, showing the intricate muscle patterns.
The sharp, serrated edge of a kitchen knife blade.
A close-up of moss sporophytes looking like an alien forest.
The crystalline structure of sugar granules spilled on a dark surface.
A micro view of a feather, showing the barbs and barbules.
The glowing filament of a lightbulb, vibrating with heat.
Camera fixed. A view of moss sporophytes.
Camera push in. Towards the nucleus of a cell.
Camera pull out. From a facet to a gemstone view.
Camera move left. Along a crack in glass.
Camera move right. Past glowing embers.
Camera orbit left. Circling a pollen grain.
Camera orbit right. Around a mosquito.
Camera pan left. Across a bird egg shell.
Camera pan right. Over a CD surface.

**Material representation.**

**Prompts:**
The iridescent surface of a soap bubble swirling with colors.
A patch of soft, white animal fur, detailed and fluffy.
Rusty iron metal surface with flaking orange and brown corrosion.
Translucent jade stone with internal cloudy patterns.
A wet cobblestone street reflecting streetlights at night.
A piece of honeycomb dripping with thick, golden honey.
The rough, porous texture of a lava rock.
Silky satin fabric flowing like liquid in a gentle breeze.
A close-up of carbon fiber weave, black and geometric.
The sparkling surface of fresh snow in sunlight.
A pool of crude oil, thick, black, and viscous.
Old, cracked leather texture on an antique armchair.
A mosaic of stained glass pieces, glowing with transmitted light.
The intricate pattern of a snake's scales.
A close-up of a woven wicker basket.
Molten lava flowing and cooling into black rock.
Camera pull out. Revealing a tire tread.
Camera move left. Along a copper pipe.
Camera move right. Past a jelly cube.
Camera orbit left. Circling a tennis ball.
Camera move right, then push in. Along a leather crack, then focusing on the texture.
Camera orbit left, transitioning into a pan right. Circling a stained glass piece, then sweeping the window.
Camera pan right, then pull out. A wide view of wet cobblestones, pulling back to the street.
Camera move left, pull out, then pan right. Starting at a sequin, revealing the dress.

### B.2.4 Fantasy and Surrealism

This category challenges the model's generalization capabilities by introducing non-Euclidean geometries and physics-defying structures.

**Prompts:**
A floating library orbiting a small star.
A city where gravity shifts between districts.
A dreamlike ruin suspended in a crystal sphere.
A surreal canyon filled with echoing light.
A floating island with an endless spiral tower.
A dream world where the ground reflects the sky.
A surreal city constructed from ancient stone faces.
A floating garden drifting through a pastel sky.
A dreamlike shoreline where water turns into glass.
A surreal city half submerged in glowing mist.
A floating citadel guarded by silent stone figures.
A dreamscape where mountains fold like paper.
A surreal field filled with hovering monoliths.
A floating city beneath an aurora-filled sky.
A dream world where light casts solid shadows.
A surreal bridge stretching across a starless void.
A floating village supported by giant birds.
A dreamlike forest where the air shimmers.
A surreal city built inside a colossal hourglass.
A floating island with inverted waterfalls.
A dreamscape where the moon rests on the horizon.
A surreal canyon illuminated by glowing symbols.
Camera move right. A dreamscape where shadows form buildings.
Camera orbit left. A surreal forest glowing softly at night.
Camera orbit right. A floating tower surrounded by glowing fog.
Camera pan left. A dream world where gravity reverses at dawn.
Camera pan right. A surreal bridge stretching across a void.
Camera push in. A floating ruin wrapped in glowing vines.
Camera pull out. A dreamscape where the sky shatters softly.
Camera move left. A surreal city illuminated by floating crystals.
Camera move right. A floating island drifting through mist.
Camera orbit left. A dream world where space folds inward.
Camera push in, then orbit left. A floating island with inverted waterfalls.
Camera pull out, then pan right. A surreal city built on clock gears.
Camera move left, then push in. A dreamscape where shadows float freely.
Camera move right, then orbit right. A floating palace drifting through fog.
Camera push in, then pan left. A surreal canyon filled with glowing glyphs.
Camera pull out, then orbit left. A floating observatory surrounded by stars.
Camera move left, then pan left. A dream world where geometry bends.
Camera move right, then pan right. A surreal forest of bioluminescent trees.
Camera push in, then orbit right. A floating city glowing at eternal dusk.
Camera pull out, then pan left. A dreamscape where the horizon curves.
Camera move left, then push in. A surreal island chained to the moon.
Camera move right, then push in. A floating ruin wrapped in glowing vines.
Camera push in, then pull out. A dream world where buildings breathe softly.
Camera orbit left, then pan left. A surreal valley bathed in violet light.
Camera orbit right, then pan right. A floating city drifting apart.

### B.2.5 Artistic Styles

This category ensures that the 3D constraints do not degrade the aesthetic diversity of the foundation model, requiring distinct stylized renderings rather than photorealism.

**Prompts:**
A campfire in snowy woods, watercolor painting
A street musician performing at night, black and white photography
A flickering neon sign in fog, cyberpunk illustration
A castle made of glass, fantasy illustration
Camera move right. A sunflower field under swirling sky, Van Gogh style
Camera orbit left. A cracked marble statue repaired with gold, kintsugi art
Camera orbit right. A glowing crystal floating in darkness, fantasy digital art
Camera push in. A girl holding a lantern, watercolor painting
Camera pull out. An infinite library interior, surrealism
Camera pan left. A snowy village at dusk, storybook illustration
Camera pan right. A futuristic skyline at night, matte painting
Camera move left. A cat sleeping on stacked books, pastel illustration
Camera move right. A mechanical bird on a branch, steampunk illustration
Camera orbit left. A floating violin, surreal illustration

## B.3 Dynamic Data Subset

This subset focuses on high-entropy scenes and deformable objects to mitigate the suppression of non-rigid dynamics during 3D alignment.

**Prompts:**
A massive waterfall cascading down a cliff with mist rising.
Camera push in. A lion roaring with its mane shaking in the wind.
Camera orbit left, then pull out. A futuristic car transforming into a robot.
Camera pan right. A cyclist sprinting towards the finish line.
Camera fixed. Raindrops splashing into a puddle, creating ripples.
Camera move left, pull out, then pan left. A spaceship launching with thrusters igniting.
A chef tossing vegetables in a flaming wok.
Camera orbit right. A tornado tearing through a wooden barn.
Camera pull out, then pan right. Hundreds of lanterns floating up into the night sky.
Camera move right. A subway train rushing past the platform.
Camera push in, then pan left. A flower blooming rapidly in time-lapse.
A glass shattering on the floor in slow motion.
Camera pan left. Soldiers marching in synchronization across a dusty field.
Camera orbit left. A breakdancer spinning on their head.
Camera move right, pull out, then pan right. A whale breaching the ocean surface.
A volcano erupting with lava flowing down the side.

# C User Study Details

To comprehensively evaluate the perceptual quality of World-R1 compared to baseline foundation models, we conducted a blind subjective user study. While quantitative metrics provide objective measures of reconstruction, human evaluation remains the gold standard for assessing generative video quality, particularly for semantic adherence and visual plausibility.

## C.1 Experimental Setup

**Methodology.** We adopted a Two-Alternative Forced Choice (2AFC) protocol. For each evaluation instance, participants were presented with two video clips played side-by-side: one generated by our World-R1 model and the other by the corresponding base model (Wan 2.1). The position of the videos (left vs. right) was randomized for every query to eliminate positional bias. The study was strictly double-blind; neither the participants nor the administrators knew which model produced which video during the evaluation phase.

**Model Comparisons.** To ensure a fair comparison, we matched models by parameter size:

- **Small Variant:** *World-R1-Small* was compared against *Wan2.1-T2V-1.3B*.
- **Large Variant:** *World-R1-Large* was compared against *Wan2.1-T2V-14B*.

**Participants.** The study involved a total of 25 participants. The pool consisted of evaluators with varying degrees of familiarity with generative AI, ensuring the results reflect both expert critique and general user preference.

## C.2 Evaluation Dataset

We curated a fixed evaluation set of **30 complex prompts** designed to stress-test the models' capabilities in world modeling. The dataset was distinct from the training set and covered a diverse range of scenarios.

## C.3 Evaluation Criteria

Participants were instructed to watch the video pairs and select the superior model based on three independent criteria. We provided specific definitions for each criterion to ensure consistent grading:

1. **Geometric Consistency:** *"Which video better maintains the structural integrity of the physical world?"*
   - **Look for:** Objects that remain solid (no morphing/vanishing), background elements that stay fixed relative to the camera movement, and correct perspective changes (parallax).
   - **Penalize:** Walls that warp, objects that float or disappear when the camera moves, and "dream-like" transitions in rigid environments.
2. **Camera Control Accuracy:** *"Which video more accurately executes the camera movement described in the text prompt?"*
   - **Look for:** Strict adherence to the directional command (e.g., if the prompt says "orbit," the camera must circle the object, not just pan or zoom).
   - **Penalize:** Errant camera drift, static cameras when motion is requested, or wrong movement directions.
3. **Overall Visual Quality:** *"Which video do you prefer overall?"*
   - **Look for:** High image fidelity, aesthetic appeal, lack of visual artifacts (blurriness, noise), and general coherence.
   - **Note:** This metric captures the trade-off between strict adherence to physics and artistic generation quality.

**Table A** User Study Results (Win Rate of World-R1 vs. Wan 2.1). The results demonstrate a decisive preference for World-R1 in tasks requiring physical reasoning and control.

| Metric | Win Rate (%) | Analysis |
|---|---|---|
| Geometric Consistency | **92%** | Superior structural stability; significant reduction in hallucination. |
| Camera Control Accuracy | **76%** | Better instruction following for complex trajectories. |
| Overall Preference | **86%** | Users prefer the aligned, consistent videos despite the constraints. |

### C.4 Detailed Results

We aggregated the responses across all participants and prompts. The Win Rate represents the percentage of instances where World-R1 was selected as superior to the baseline *Wan 2.1*.

The results indicate that while the baseline models are capable of generating visually pleasing individual frames, they frequently fail to maintain 3D consistency over time. World-R1's dominant performance in Geometric Consistency (92%) validates the effectiveness of our RL-based alignment strategy in teaching the model latent 3D constraints.

### C.5 Metric Validation Study

To further verify that our automatic 3D-consistency metric reflects human perception rather than artifacts of the reconstruction pipeline, we conducted an additional metric-validation user study. This study involved 20 participants and 30 randomized video pairs. For each pair, participants selected the video with stronger perceived 3D consistency, and we compared the majority human preference with the ranking induced by our automatic metric. As shown in Table B, human preference aligns with the metric ranking in 91.17% of comparisons, confirming that the metric is a reliable perceptual proxy for geometric consistency.

**Table B** Metric-validation user study. Agreement measures how often majority human preference on 3D consistency matches the automatic metric ranking.

| Study | Participants | Video Pairs | Agreement ↑ |
|---|---|---|---|
| 3D-consistency metric validation | 20 | 30 | **91.17%** |

## D Additional Experimental Analyses

This section provides additional experiments motivated by reviewer feedback. We place these analyses in the appendix to keep the main paper focused while making the empirical validation more comprehensive.

### D.1 Camera-Control Accuracy

We further evaluate camera-control accuracy using RotErr, TransErr, and CamMC under a reference-video trajectory protocol. Despite being trained as a post-training alignment method rather than a dedicated camera-control architecture, World-R1 achieves camera-control accuracy competitive with specialized methods while preserving stronger general video quality.

### D.2 Reconstruction-Independent Multi-View Consistency

Our main 3D-consistency metric uses a shared 3DGS reconstruction pipeline for all compared methods, which ensures a fair comparison but may still inherit biases from the reconstruction

**Table C** Camera-control evaluation. Lower RotErr, TransErr, and CamMC are better.

| Method | RotErr ↓ | TransErr ↓ | CamMC ↓ |
|---|---|---|---|
| ReCamMaster [37] | 1.53 | 3.12 | 4.17 |
| TrajectoryCrafter [60] | 3.08 | 7.46 | 10.22 |
| CamCloneMaster [61] | 1.36 | 2.02 | 3.05 |
| Wan2.1-T2V-1.3B [3] | 9.29 | 62.94 | 66.21 |
| Wan2.1-T2V-14B [3] | 17.01 | 60.90 | 70.55 |
| **World-R1-Small** | 1.50 | 2.76 | 3.39 |
| **World-R1-Large** | **1.21** | **1.30** | **2.95** |

model. To verify that the improvements do not merely come from the reconstruction process, we additionally evaluate Multi-View Consistency Score (MVCS) following GeoVideo [62], a reconstruction-independent metric that directly measures cross-view agreement from generated videos. As shown in Table D, World-R1 improves MVCS for both the small and large backbones, confirming that the gains stem from the generated videos rather than artifacts of 3DGS optimization.

**Table D** Reconstruction-independent multi-view consistency evaluation. Higher MVCS indicates stronger cross-view agreement.

| Method | MVCS ↑ |
|---|---|
| Wan2.1-T2V-1.3B [3] | 0.974 |
| **World-R1-Small (Ours)** | **0.989** |
| Wan2.1-T2V-14B [3] | 0.963 |
| **World-R1-Large (Ours)** | **0.993** |

## D.3 Dataset Scaling

We evaluate the effect of increasing the size of the pure text post-training dataset. Table E shows a consistent positive trend from 1K to 3K prompts across both 3D consistency and general video quality. The result indicates that World-R1 is data-efficient, already producing strong gains with 3K prompts, and suggests that further scaling of prompt generation could yield additional improvements.

**Table E** Dataset scaling study on World-R1-Small. Increasing prompt data improves both geometric consistency and general video quality.

| Data Size | PSNR ↑ | SSIM ↑ | LPIPS ↓ | VBench AVG ↑ |
|---|---|---|---|---|
| 1K | 25.82 | 0.812 | 0.258 | 83.23 |
| 2K | 26.54 | 0.839 | 0.223 | 84.76 |
| 3K | **27.63** | **0.858** | **0.201** | **85.21** |

## D.4 Long-Video Generalization

Although training is conducted on short clips, World-R1 generalizes to longer generations. We evaluate 121-frame videos in Table F. World-R1-Large substantially improves all 3D-consistency metrics over the Wan2.1-T2V-14B backbone, suggesting that the learned geometric alignment transfers beyond the training horizon.

**Table F** Long-video evaluation on 121-frame generations.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| Wan2.1-T2V-14B [3] | 18.32 | 0.558 | 0.534 |
| **World-R1-Large (Ours)** | **26.32** | **0.828** | **0.257** |

## D.5 Scene-Complexity Breakdown

To evaluate challenging settings more systematically, we categorize the test prompts by scene type and report per-category results in Table G. World-R1-Small consistently improves over the Wan2.1-T2V-1.3B backbone across static scenes, single-object dynamics, multi-object dynamics, non-rigid motion, and selected long-horizon dynamics. The largest absolute challenge remains long-horizon and non-rigid generation, where final quality is partly limited by the base model's compositional and motion-generation capacity; nevertheless, the improvements show that RL-based 3D alignment remains effective in these regimes.

**Table G** Per-category results by scene complexity. "N" denotes the percentage of prompts in each category. Long-horizon dynamics are selected from the preceding categories.

| Scene Type | N | Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | MVCS ↑ |
|---|---|---|---|---|---|---|
| Static Scene | 30.11% | Wan2.1-1.3B | 20.14 | 0.632 | 0.389 | 0.981 |
| Static Scene | 30.11% | **World-R1-Small** | **30.52** | **0.912** | **0.142** | **0.994** |
| Single-obj Dynamic | 29.03% | Wan2.1-1.3B | 17.86 | 0.563 | 0.452 | 0.976 |
| Single-obj Dynamic | 29.03% | **World-R1-Small** | **28.17** | **0.869** | **0.189** | **0.991** |
| Multi-obj Dynamic | 21.51% | Wan2.1-1.3B | 15.23 | 0.487 | 0.528 | 0.968 |
| Multi-obj Dynamic | 21.51% | **World-R1-Small** | **25.41** | **0.812** | **0.248** | **0.985** |
| Non-rigid Motion | 19.35% | Wan2.1-1.3B | 14.58 | 0.462 | 0.548 | 0.965 |
| Non-rigid Motion | 19.35% | **World-R1-Small** | **24.73** | **0.793** | **0.267** | **0.982** |
| Long-horizon Dynamics | 12.89% | Wan2.1-1.3B | 12.53 | 0.382 | 0.683 | 0.951 |
| Long-horizon Dynamics | 12.89% | **World-R1-Small** | **23.59** | **0.781** | **0.299** | **0.974** |

## D.6 Comparison with 3D-Aware Generation Methods

We further consolidate comparisons with publicly available 3D-aware video generation methods. These methods fall into two categories: camera-control methods that introduce explicit control modules, and 3D-conditioned methods that rely on architectural modifications or image-to-video pipelines. In contrast, World-R1 is a text-to-video post-training framework requiring no inference-time architectural changes. As shown in Table H, World-R1 achieves the strongest 3D-consistency metrics while preserving competitive general video quality. These results indicate that our post-training paradigm is complementary to architecture-level 3D-aware designs and can potentially be applied on top of them.

## D.7 Reward Hacking Analysis

A potential failure mode of RL-based alignment is reward hacking: if the model optimizes only a narrow 3D metric, it may learn degenerate shortcuts such as producing near-static clips that are easy to reconstruct but do not follow the desired camera motion or preserve dynamic content. Our design mitigates this issue in three ways. First, the composite reward combines meta-view plausibility, reconstruction fidelity, trajectory alignment, and general video-generation quality, so no single metric can dominate the optimization. Second, the trajectory term penalizes static outputs when a non-trivial camera trajectory is requested. Third, periodic decoupled training on dynamic prompts prevents the model from collapsing into overly rigid generation.

**Table H** Consolidated comparison with 3D-aware and camera-control generation methods. The 3D-conditioned baselines are evaluated in their image-to-video setting, whereas World-R1 is evaluated as a text-to-video model on general scenes.

| Type | Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | MVCS ↑ | Aesthetic ↑ | BG Cons. ↑ | Subject Cons. ↑ | Motion Smooth. ↑ |
|---|---|---|---|---|---|---|---|---|---|
| 3D-Cond. | ViewCrafter [63] | 23.15 | 0.724 | 0.291 | 0.979 | 55.52 | 92.09 | 94.25 | 97.86 |
| 3D-Cond. | Voyager [8] | 21.38 | 0.678 | 0.334 | 0.975 | 49.80 | 92.31 | 91.55 | 99.39 |
| 3D-Cond. | FlashWorld [64] | 22.46 | 0.702 | 0.312 | 0.977 | 53.72 | 91.88 | 94.44 | 98.81 |
| 3D-Cond. | VerseCrafter [65] | 23.82 | 0.748 | 0.268 | 0.981 | 54.78 | 94.88 | 95.55 | 97.62 |
| Cam. Ctrl. | GCD [56] | 18.26 | 0.582 | 0.438 | 0.966 | 38.21 | 92.00 | 88.94 | 98.37 |
| Cam. Ctrl. | Traj.-Attn. [57] | 18.87 | 0.598 | 0.421 | 0.969 | 38.50 | 92.83 | 90.60 | 98.21 |
| Cam. Ctrl. | DAS [58] | 19.42 | 0.618 | 0.398 | 0.971 | 39.86 | 92.03 | 90.34 | 99.14 |
| Cam. Ctrl. | ReCamMaster [37] | 20.58 | 0.653 | 0.368 | 0.975 | 42.70 | 93.83 | 92.05 | 99.28 |
| Foundation | Wan2.1-1.3B [3] | 17.40 | 0.550 | 0.467 | 0.974 | 62.43 | 97.29 | 96.34 | 97.44 |
| Ours | **World-R1-Small** | **27.63** | **0.858** | **0.201** | **0.989** | **65.74** | 96.67 | **97.58** | 98.55 |

Table I reports component-wise reward ablations. Removing any reward component degrades performance, and the full pipeline achieves the best overall trade-off across 3D consistency and VBench quality. In particular, removing $\mathcal{S}_{traj}$ weakens trajectory adherence, while removing $\mathcal{S}_{meta}$ or $\mathcal{S}_{recon}$ reduces geometric consistency. Table J further analyzes the core training and conditioning components together with the decomposed reward signals. Removing periodic decoupled training can improve reconstruction-style scores by pushing the model toward overly rigid or near-static solutions, but it degrades VBench quality. Removing the 3D-aware reward preserves general video quality but fails to regularize geometry. These results support that the reward components act as complementary safeguards against reward exploitation rather than redundant terms.

**Table I** Reward component ablation on World-R1-Small. Each component contributes to the final balance between geometric consistency and general video quality.

| Variant | PSNR ↑ | SSIM ↑ | LPIPS ↓ | VBench AVG ↑ |
|---|---|---|---|---|
| Full pipeline (World-R1-Small) | **27.63** | **0.858** | **0.201** | **85.21** |
| w/o $\mathcal{S}_{meta}$ | 26.91 | 0.841 | 0.218 | 83.67 |
| w/o $\mathcal{S}_{recon}$ | 25.14 | 0.798 | 0.271 | 84.35 |
| w/o $\mathcal{S}_{traj}$ | 26.27 | 0.829 | 0.237 | 84.53 |

**Table J** Training and conditioning ablation with decomposed reward scores. Dashes indicate disabled reward terms.

| Variant | PSNR ↑ | SSIM ↑ | LPIPS ↓ | VBench AVG ↑ | $\mathcal{S}_{recon}$ ↑ | $\mathcal{S}_{traj}$ ↑ | $\mathcal{S}_{meta}$ ↑ | $\mathcal{S}_{gen}$ ↑ |
|---|---|---|---|---|---|---|---|---|
| Full | **27.63** | 0.858 | 0.201 | **85.21** | 0.342 | 0.296 | **0.307** | **0.37** |
| w/o noise wrapping | 24.46 | 0.745 | 0.298 | 76.39 | 0.312 | 0.213 | 0.275 | -0.42 |
| w/o periodic decoupled training | 27.89 | **0.898** | **0.192** | 82.64 | 0.348 | **0.310** | 0.298 | 0.18 |
| w/o 3D-aware reward | 18.93 | 0.502 | 0.496 | 84.96 | – | – | – | 0.33 |
| w/o general reward | 27.57 | 0.849 | 0.206 | 83.44 | **0.388** | 0.231 | 0.305 | – |

# E More Visualizations

For video results, please refer to the video in supplementary material.

**Dynamic video results.** A core challenge in aligning video generation models with 3D constraints is the potential suppression of non-rigid dynamics. However, thanks to our periodic decoupled training strategy and the utilization of the dynamic data subset, World-R1 successfully disentangles camera motion from object motion. As shown in Figure B, our model retains the high-entropy generative capabilities of the foundation model. These results confirm that World-R1 as a robust world simulator, enforcing geometric consistency on the static environment while permitting correct physical dynamics for non-rigid objects.

**3D consistency via 3D reconstruction.** To explicitly validate the geometric integrity of the synthesized worlds, we conduct a reconstruction-based analysis. We compare the original videos generated by *World-R1* with their corresponding re-renderings derived from 3DGS representation (as detailed in the reward formulation in Section A.1). Since 3DGS optimization assumes a static scene observed from moving viewpoints, it is highly sensitive to geometric inconsistencies. As illustrated in Figure D and Figure C, the difference in reconstruction quality serves as a direct proxy for the underlying video consistency. The high visual fidelity of the 3DGS re-renderings closely matching the original generated content, demonstrates that World-R1 produces 3D consistent video sequences that adhere to 3D constraints, effectively minimizing geometric hallucinations.

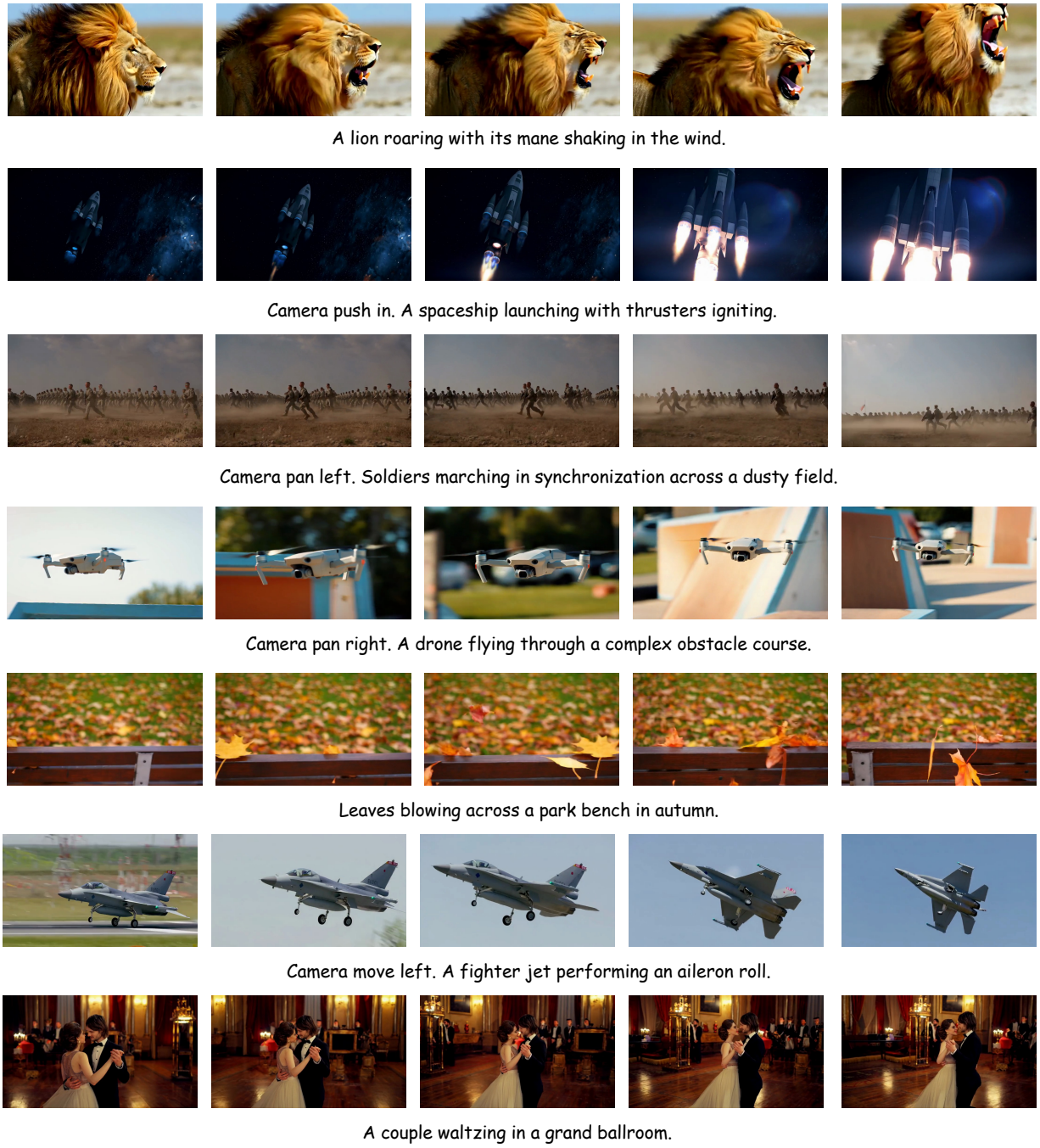


**Figure B** **Qualitative results on dynamic scenes.** We visualize samples from World-R1 on prompts requiring high temporal dynamics. Despite being trained with strict 3D consistency rewards, the model preserves the natural motion of non-rigid elements, demonstrating the effectiveness of our periodic decoupled training strategy.

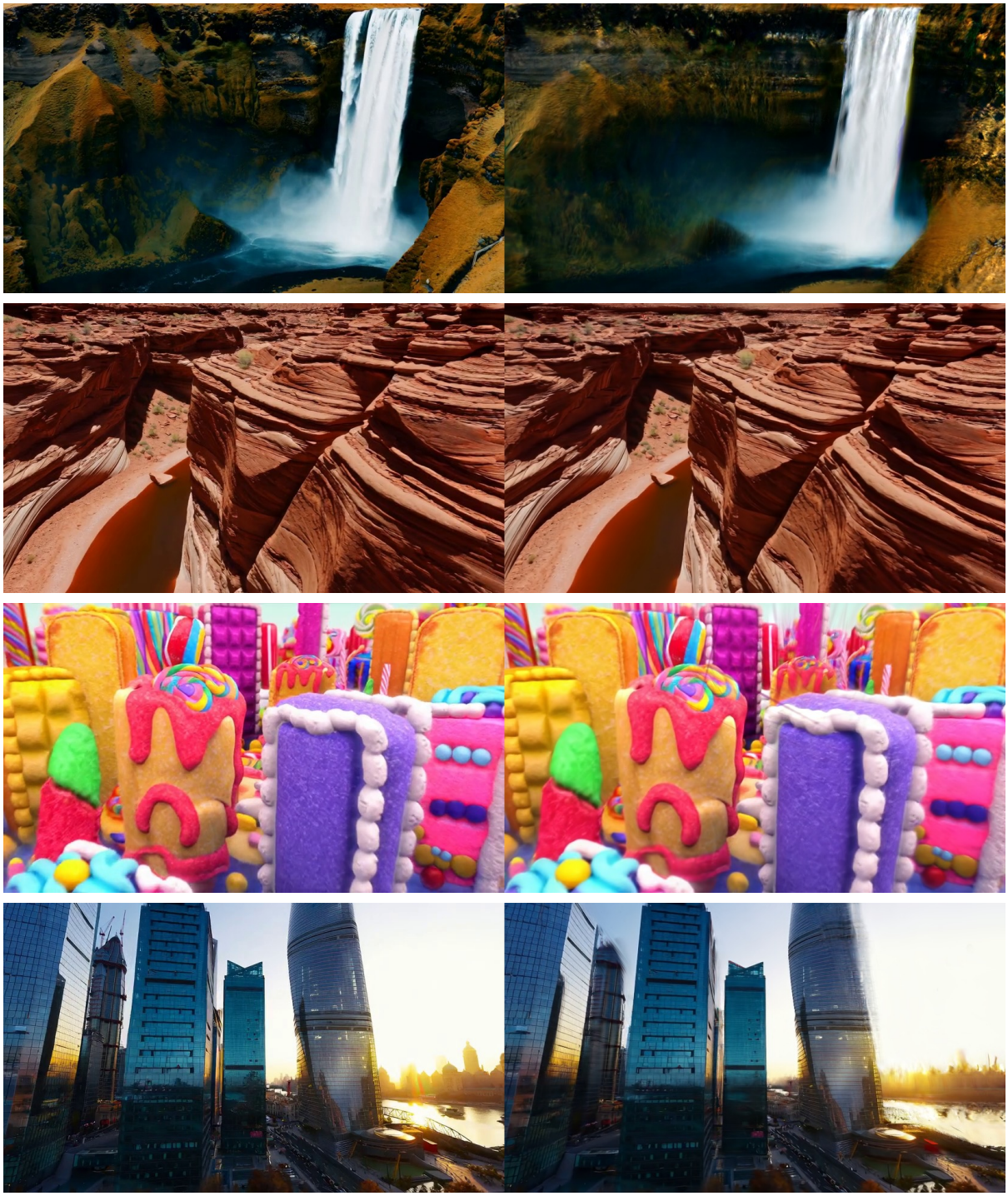


**Figure C High-Fidelity Reconstruction from World-R1.** The strict 3D consistency preserved by World-R1 enables the recovery of a dense and accurate 3D scene representation, free from the floaters and distortion observed in baseline methods.

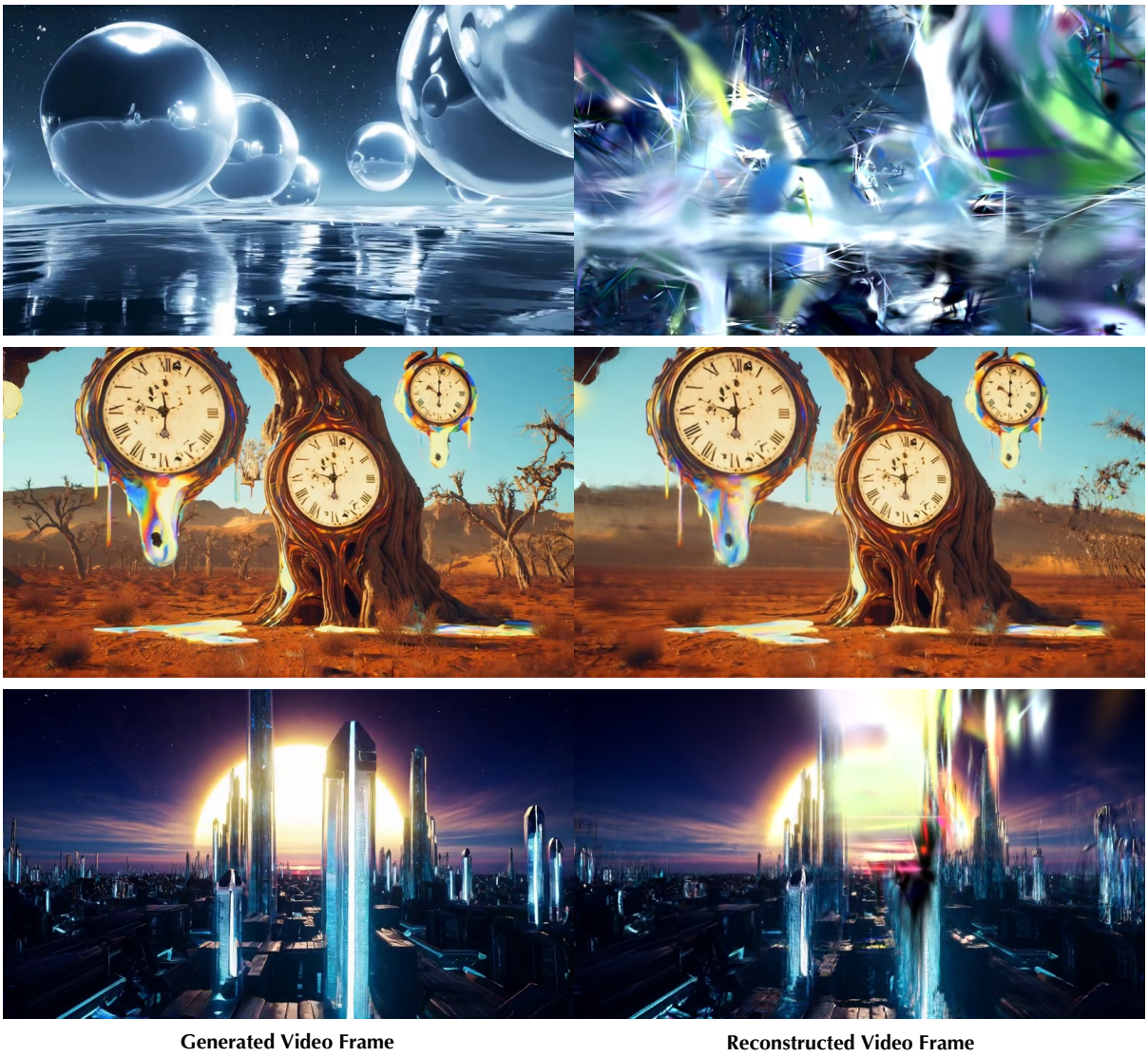


**Figure D Reconstruction Failure due to Geometric Inconsistency.** Visualizations of some reconstruction results derived from the videos of Wan 2.1 [3]. Due to temporal warping and object morphing in the source video, the 3D optimization fails to form a coherent structure, resulting in significant noise, sparse point clouds, and visual artifacts.