EFFICIENT POST-TRAINING FOR 3D-CONSISTENT VIDEO GENERATION VIA OPTIMAL TRANSPORT REWARD BALANCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) can improve the 3D awareness of video diffusion models without architectural modifications. However, existing methods face three challenges: reconstruction rewards favor limited camera motion, heterogeneous reward distributions cause imbalanced optimization, and policy updates over many denoising steps are computationally expensive. We propose 3D-Balanced-R1 to address these issues. First, a visibility score measures inter-frame overlap and reweights the reconstruction reward according to reconstruction difficulty. Second, optimal transport aligns different rewards to a common Gaussian distribution, reducing domination by individual rewards or outliers. Third, we update the policy at only two denoising steps, using an early step for camera control and a late step for video quality and 3D consistency. On Wan2.1-1.3B, 3D-Balanced-R1 achieves a $3.53\times$ training speedup over World-R1. With the same number of training epochs and comparable video quality, 3D-Balanced-R1 improves the pretrained model's 3D score by 9.6%, compared with 6.3% for World-R1. Under a similar training-time budget, the gain increases to 14.3%, or $2.3\times$ that of World-R1. Similar improvements are also observed on the larger Wan2.2-5B model.

1 INTRODUCTION

Video diffusion models Blattmann et al. (2023b;a); Ho et al. (2022); Ma et al. (2025a); Singer et al. (2023); Yang et al. (2025); Wan et al. (2025) can now synthesize photorealistic and temporally coherent videos from text, yet training on 2D videos alone does not explicitly enforce an understanding of the underlying 3D scene. Recent work therefore leverages geometry foundation models (GFMs) Wang et al. (2024a; 2025); Lin et al. (2026); Wang et al. (2026a) to provide 3D supervision for video diffusion models. One line of work applies supervised fine-tuning (SFT) on static, 3D-consistent data Gao* et al. (2024); Liu et al. (2024); Wu et al. (2025); Zhu et al. (2026); Wu et al. (2026a), which can bias generation toward static scenes and weaken the generalization of the pretrained model.

More recent approaches Wang et al. (2026b); An et al. (2026) instead use Flow-GRPO Liu et al. (2026a) with GFM-derived rewards to improve 3D awareness while largely preserving general video-generation capabilities. However, RL post-training remains computationally expensive, and its effectiveness is sensitive to the design of 3D-aware rewards. World-R1 Wang et al. (2026b) trains for approximately 120 hours on eight GPUs, yet Table 1 shows that it improves trajectory alignment by only 4.3% over Wan2.1-1.3B. We therefore ask *what limits the effectiveness and efficiency of RL post-training for 3D-aware video generation?* Based on our analysis, we propose 3D-Balanced-R1, which improves overall 3D-aware generation performance while substantially reducing memory and computational cost. On the eight-GPU setting used for the efficiency comparison, 3D-Balanced-R1 achieves a $3.53\times$ training speedup.

These RL-based approaches use reconstruction consistency and camera-trajectory scores Wang et al. (2026b); An et al. (2026) to improve 3D awareness, but the two objectives can conflict. Small viewpoint changes produce high inter-frame overlap and are easier to reconstruct, so reconstruction scores based on reprojection error or rendering quality can favor limited camera motion. Larger camera motion is harder to reconstruct even when it better follows the prompt. Moreover, the trajectory

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

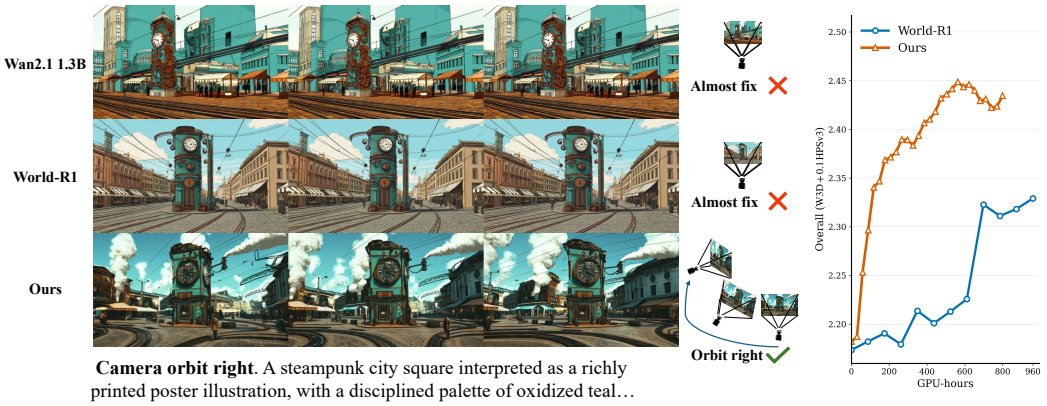


Figure 1: **Comparison with World-R1. Left:** Our method better follows camera-trajectory prompts while maintaining 3D consistency. **Right:** Our method converges much faster during RL post-training with substantially lower computational cost.

score alone does not fully resolve this conflict because normalized camera poses do not directly capture the magnitude of viewpoint changes. We therefore measure inter-frame overlap with a visibility score and use it to reweight the reconstruction reward. This accounts for reconstruction difficulty across different viewpoint changes and reduces the reward's bias toward nearly static cameras.

Beyond this reward conflict, we observe substantial distribution mismatch across rewards, as illustrated in the bottom-right of Fig. 3. Although Flow-GRPO normalizes the advantage of each reward Shao et al. (2024), this normalization does not eliminate differences in their underlying distributions. Rewards with larger variance or faster changes can therefore dominate the combined advantage. This behavior is particularly evident during the early stage of World-R1 training Wang et al. (2026b), where HPSv3 Ma et al. (2025b) dominates the reward signal and largely determines whether a sampled video receives a positive or negative advantage. Consequently, the policy is initially optimized primarily toward HPSv3 rather than improving all reward objectives jointly, while reconstruction and trajectory performance can even regress. This imbalance makes joint optimization less effective.

To address this reward imbalance, we introduce optimal transport to align the distribution of each reward to a standard Gaussian, followed by 3σ clipping. Mapping heterogeneous reward distributions to a common reference distribution reduces differences in scale and tail behavior, preventing a single reward or extreme rollout from dominating the combined advantage. Consequently, the resulting advantage better reflects balanced improvement across multiple objectives rather than an extreme value from one reward. This distribution alignment stabilizes multi-reward optimization and leads to stronger final performance.

Finally, previous methods follow the standard Flow-GRPO pipeline, which stores multiple denoising steps during rollout generation for subsequent online policy updates. Inspired by Flash-GRPO He et al. (2026), we retain only two denoising steps: an early step that strengthens text-guided camera control, and a late step that maintains video quality and fine-grained 3D consistency. This sparse-step update substantially reduces both memory consumption and policy-update cost, improving overall training efficiency.

In this paper, we identify three factors that limit effective and efficient Flow-GRPO training for 3D-aware video generation. Our full Wan2.1-1.3B run uses eight H100 GPUs for approximately 100 hours. It improves W3D from 1.235 for the pretrained backbone and 1.313 for World-R1 to 1.412, while increasing HPSv3 from 9.389 to 10.229. Under an equal-epoch comparison, 3D-Balanced-R1 also achieves a higher W3D score than World-R1 (1.353 versus 1.313). The higher W3D score relative to World-R1 extends to Wan2.2-5B, where 3D-Balanced-R1 reaches 1.365 compared with 1.240 for World-R1. Our contributions are summarized below.

- We introduce a visibility score based on inter-frame overlap to account for reconstruction difficulty. Reweighting the reconstruction reward with this score reduces its preference for static-camera videos and mitigates its conflict with the trajectory reward.

- We use optimal transport with three-sigma clipping to place heterogeneous rewards on a comparable scale and limit the influence of outliers during policy optimization.
- We reduce the cost of online policy updates by retaining two denoising transitions, one early step for camera motion and one late step for video quality and fine-grained 3D consistency.
- We validate 3D-Balanced-R1 on Wan2.1-1.3B and Wan2.2-5B, obtaining higher W3D scores while maintaining video quality and substantially reducing policy-update time.

2 Related Work

3D-conditioned video generation. Recent diffusion Transformer (DiT) Peebles & Xie (2023)-based video generation models Yang et al. (2025); Kong et al. (2024); Wan et al. (2025) have demonstrated strong capabilities in generating high-quality and temporally coherent videos. To enable explicit control over viewpoint changes, CameraCtrl, MotionCtrl, and ReCamMaster He et al. (2024); Wang et al. (2024b); Bai et al. (2025) condition video generation on camera poses to synthesize videos along specified camera trajectories, while CAT3D Gao* et al. (2024) and SV3D Voleti et al. (2024) leverage multi-view conditioning to generate geometrically consistent novel views for 3D reconstruction. More explicit geometric conditioning has also been explored. GEN3C Ren et al. (2025) and 3DGS-Enhancer Liu et al. (2024) condition video synthesis on rendered images from renderable 3D representations, while recent world models Huang et al. (2025); Shen et al. (2026); Wu et al. (2026b); Zhu et al. (2026) leverage explicit geometry, such as point clouds or persistent spatial memory, to support consistent scene exploration and world generation. In contrast to directly conditioning the generator on camera poses or explicit geometric representations, Go-with-the-Flow Burgert et al. (2025) controls video generation by warping the input noise, providing geometry-aware control without introducing additional camera conditions to the model. However, these methods often require supervised fine-tuning (SFT) on costly 3D data, which may degrade the generalization from large-scale pretrained video generation models. In this work, we use an RL-based post-training framework to improve 3D awareness while preserving general video-generation capabilities.

Video generation with 3D awareness. Instead of relying on explicit 3D conditions during generation, another line of work improves the 3D awareness of video models through geometry-aware training objectives. Along this direction, Geometry Forcing Wu et al. (2026a) and VideoWeave Xiang et al. (2026) align video representations with geometry-aware features or latents derived from geometry foundation models Wang et al. (2025). More recently, reinforcement learning has emerged as an effective way to improve the 3D awareness of pretrained video diffusion models while largely preserving their general generation capability. Instead of retraining the model with additional 3D supervision, these methods use rewards derived from geometry foundation models Wang et al. (2024a; 2025); Lin et al. (2026) to guide post-training. World-R1 Wang et al. (2026b) combines video quality rewards with reconstruction-based geometric rewards, using Depth Anything 3 Lin et al. (2026) to predict novel views and comparing them with the corresponding generated frames to measure 3D consistency, while VGGRPO An et al. (2026) introduces latent-space geometric rewards to improve camera smoothness and reprojection consistency with lower VAE-decoding cost. However, 3D-consistency rewards can favor videos with limited viewpoint changes, as they tend to achieve higher similarity-based scores. To address this issue, we introduce a visibility reward based on inter-frame overlap to better account for reconstruction difficulty.

Reward-Based Post-Training of Diffusion Models. DDPO Black et al. (2024) and DPOK Fan et al. (2023) develop policy-gradient methods for diffusion-model fine-tuning, while InstructVideo Yuan et al. (2024) extends reward-based optimization to video generation. More recently, Flow-GRPO Liu et al. (2026a) adapts GRPO Shao et al. (2024) to flow-matching models, and DanceGRPO Xue et al. (2025) extends group-relative optimization to image and video generation. DiffusionNFT Zheng et al. (2026) and Flash-GRPO He et al. (2026) focus on improving training efficiency, while GDPO Liu et al. (2026b), MARBLE Zhao et al. (2026), and RDPO Bai et al. (2026) study the aggregation of multiple reward signals with different scales and distributions. Our method builds on these developments for 3D-aware video diffusion models, with a focus on efficient policy updates and robust aggregation of heterogeneous geometric rewards.

3 PRELIMINARIES

Group Relative Policy Optimization for rectified flow. For rectified flow Liu et al. (2022), a clean video latent $\boldsymbol{x}_{\text{data}}$ and Gaussian noise $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ are connected by the linear path

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_{\text{data}} + t\boldsymbol{z}, \qquad t \in [0,1], \tag{1}$$

and the velocity model $v_\theta(\boldsymbol{x}_t, t, c)$, conditioned on a text prompt c, is trained with the velocity target $\boldsymbol{z} - \boldsymbol{x}_{\text{data}}$. During denoising, the model follows the ordinary differential equation (ODE)

$$\mathrm{d}\boldsymbol{x}_t = v_\theta(\boldsymbol{x}_t, t, c)\,\mathrm{d}t, \tag{2}$$

from Gaussian noise at $t = 1$ to a clean video latent at $t = 0$.

However, GRPO requires stochastic per-step likelihoods for policy optimization, whereas ODE denoising is deterministic. Flow-GRPO Liu et al. (2026a) therefore converts the ODE into a stochastic differential equation (SDE) with the same marginal distributions. With T denoising steps and $\Delta t = 1/T$, Euler–Maruyama discretization gives

$$\boldsymbol{x}_{t-1} = \boldsymbol{x}_t - \left[v_\theta(\boldsymbol{x}_t, t, c) + \frac{\sigma_t^2}{2t}\left(\boldsymbol{x}_t + (1-t)v_\theta(\boldsymbol{x}_t, t, c)\right)\right]\Delta t + \sigma_t\sqrt{\Delta t}\,\boldsymbol{\epsilon}_t, \tag{3}$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $\sigma_t = a\sqrt{t/(1-t)}$ controls the sampling stochasticity. Eq. equation 3 therefore defines a Gaussian policy $\pi_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, c)$ with a tractable per-step likelihood.

Policy optimization. For each text condition c, Flow-GRPO samples a group of G denoising trajectories and evaluates each generated video with a scalar reward R^i. The group-relative advantage is computed as

$$A^i = \frac{R^i - \text{mean}\big(\{R^j\}_{j=1}^G\big)}{\text{std}\big(\{R^j\}_{j=1}^G\big)}, \tag{4}$$

and is shared across all denoising steps of the same trajectory. For each step, the policy ratio is

$$\rho_t^i(\theta) = \frac{\pi_\theta(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, c)}{\pi_{\theta_{\text{old}}}(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, c)}. \tag{5}$$

The model is optimized with the clipped GRPO objective

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{GT}\sum_{i=1}^{G}\sum_{t=1}^{T}\left[\min\big(\rho_t^i(\theta)A^i, \text{clip}\left(\rho_t^i(\theta), 1-\varepsilon, 1+\varepsilon\right)A^i\big) - \beta D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\right]\right], \tag{6}$$

where $\pi_{\theta_{\text{old}}}$ is the policy used to sample the trajectories, π_{ref} is the frozen pretrained policy, ε is the clipping threshold, and β controls the KL regularization.

Geometry Foundation Models (GFMs). GFMs Wang et al. (2024a; 2026a); Lin et al. (2026) are feed-forward networks that directly predict scene geometry from images without test-time optimization or known camera parameters. Given N RGB frames from the same scene, a GFM Φ predicts a geometric representation for each frame:

$$\{\boldsymbol{O}_i\}_{i=1}^N = \Phi\big(\{\boldsymbol{I}_i\}_{i=1}^N\big), \qquad \boldsymbol{O}_i = \{\boldsymbol{C}_i, \boldsymbol{D}_i, \boldsymbol{P}_i\}, \tag{7}$$

where $\boldsymbol{I}_i$ denotes the i-th input frame, and $\boldsymbol{C}_i$, $\boldsymbol{D}_i$, and $\boldsymbol{P}_i$ denote its camera pose, depth map, and 3D point map, respectively. These predictions are expressed in a shared coordinate frame, allowing a single forward pass to recover a consistent 3D representation of the input video.

Some GFMs also support novel-view synthesis by using an additional Gaussian head to predict 3D Gaussian primitives from the same features, enabling the reconstructed scene to be rendered from viewpoints not observed in the input frames. In this work, we use this capability of Depth Anything 3 (DA3) Lin et al. (2026) to compute our reconstruction score.

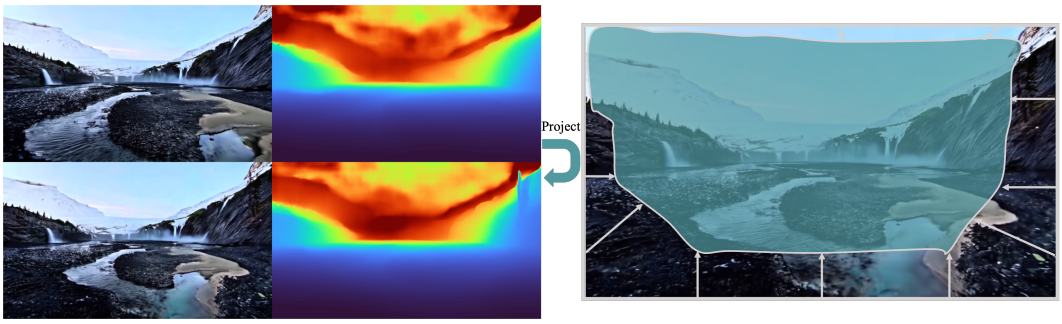


Figure 2: Visibility is computed by back-projecting depth map into 3D and reprojecting them between selected frames. Blue denotes the visible overlapping region; a larger invisible region yields a higher visibility score and indicates greater reconstruction difficulty.

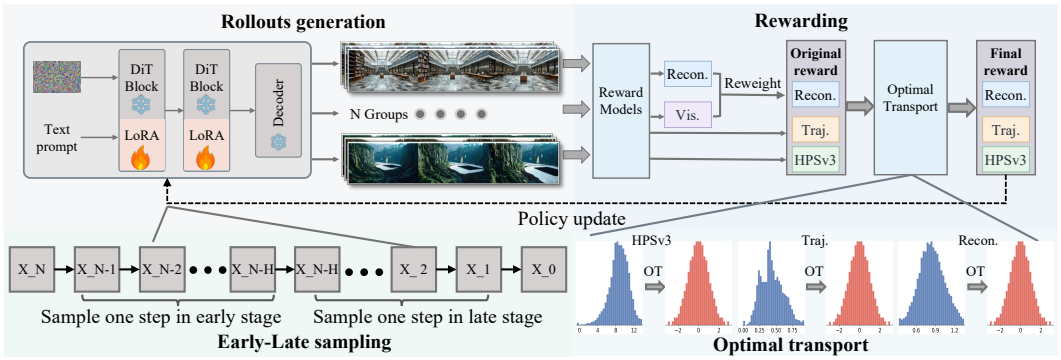


Figure 3: **Overview of 3D-Balanced-R1.** For each prompt, the LoRA-tuned video diffusion model generates a group of rollout videos. The reconstruction reward is reweighted by visibility, and optimal transport (OT) maps each reward to a standard Gaussian (bottom right). The policy is then updated on only two denoising steps, one sampled from the early stage and one from the late stage.

4 METHOD

In this section, we detail 3D-Balanced-R1, a framework based on Flow-GRPO for improving the 3D awareness of video diffusion models. As shown in Fig. 3, 3D-Balanced-R1 consists of three key components. First, we introduce the *visibility score*, which accounts for variations in 3D reconstruction difficulty by measuring cross-view visibility changes (Sec. 4.1). Second, we present *reward balancing* to mitigate the scale and distribution mismatches among different reward signals (Sec. 4.2). Finally, we develop an efficient training framework based on Flow-GRPO to substantially reduce the training cost. Together, these components enable more stable and efficient optimization, allowing 3D-Balanced-R1 to achieve stronger performance with considerably less training time.

4.1 VISIBILITY SCORE FOR 3D RECONSTRUCTION DIFFICULTY

Reconstruction-based rewards measure 3D consistency through the similarity between reconstructed and generated views. However, the reconstruction score is also affected by reconstruction difficulty. Videos with limited viewpoint changes usually have high inter-frame overlap and can therefore obtain high reconstruction scores even without strong 3D awareness. This can bias the model toward small camera motion and conflict with trajectory rewards that encourage larger viewpoint changes. To address this issue, we introduce a *visibility score* that uses inter-frame overlap as a proxy for reconstruction difficulty. Unlike estimated camera-motion magnitude, visibility is not affected by the arbitrary scale of camera translations predicted by GFMs.

Given the GFM outputs in Eq. equation 7, we uniformly sample $M = 5$ frames $\{\boldsymbol{I}_i\}_{i=0}^{M-1}$ with their corresponding geometric outputs. We compute visibility in both directions between consecutive sampled frames, with $i^+ = i + 1$. Since the predicted point maps share the same coordinate frame, we project $\boldsymbol{P}_i$ onto the next frame using its predicted camera parameters. Let $\mathcal{U}_{i^+}$ denote the valid

positive-depth pixels in the target frame, and let $\hat{\boldsymbol{D}}_{i^+}^{i\to i^+}(\boldsymbol{p})$ denote the reprojected depth at pixel $\boldsymbol{p}$. The visibility mask is defined as

$$V_{i\to i^+}(\boldsymbol{p}) = \begin{cases} 1, & \boldsymbol{p} \text{ has a valid reprojection and } \hat{\boldsymbol{D}}_{i^+}^{i\to i^+}(\boldsymbol{p}) - \boldsymbol{D}_{i^+}(\boldsymbol{p}) \le \tau_{\text{vis}}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

where τ_{vis} is the depth tolerance. The reverse mask $V_{i^+\to i}$ is defined analogously by swapping the source and target frames, with $\mathcal{U}_i$ denoting the target's valid positive-depth pixels. We average both directions over the $M-1$ consecutive frame pairs:

$$S_{\text{vis}} = 1 - \frac{1}{2(M-1)} \sum_{i=0}^{M-2} \left(\frac{1}{|\mathcal{U}_{i^+}|} \sum_{\boldsymbol{p}\in\mathcal{U}_{i^+}} V_{i\to i^+}(\boldsymbol{p}) + \frac{1}{|\mathcal{U}_i|} \sum_{\boldsymbol{p}\in\mathcal{U}_i} V_{i^+\to i}(\boldsymbol{p}) \right) \tag{9}$$

A larger S_{vis} indicates lower inter-frame overlap and therefore greater reconstruction difficulty. We use it to reweight the original reconstruction reward $r_{\text{ori}}^{\text{r}}$ as

$$r^{\text{r}} = r_{\text{ori}}^{\text{r}} \left(1 + \lambda_{\text{vis}} S_{\text{vis}}\right), \tag{10}$$

where λ_{vis} controls the reweighting strength. We clip S_{vis} at 0.2 to prevent reward exploitation from excessively large viewpoint changes. Thus, lower inter-frame overlap yields a larger reconstruction reward for the same reconstruction score.

4.2 Optimal Transport for Multi-Reward Balancing

The reward components used by GRPO, including visual quality $\mathbf{r}_{\boldsymbol{v}}$, reconstruction $\mathbf{r}_{\boldsymbol{r}}$, and trajectory alignment $\mathbf{r}_{\boldsymbol{t}}$, have substantially different scales and distributions. Directly combining them can allow components with larger variance or heavier tails to dominate the advantage, even when all components are assigned equal weights. To address this problem, we independently align each reward distribution to a common Gaussian scale using one-dimensional optimal transport before computing the GRPO advantage.

Let r_i^k denote the k-th reward component of the i-th rollout in a gathered batch of size B, with $i = 1, \ldots, B$ and $k \in \{\text{v}, \text{r}, \text{t}\}$. We map them to follow $\mathcal{N}(0,1)$ via 1-D optimal transport. We first sort the samples to obtain order statistics $r_{(1)}^k \le r_{(2)}^k \le \cdots \le r_{(B)}^k$ and assign each the empirical CDF value $F_{r^k}(r_{(j)}^k) = \frac{j-1/2}{B}$, where the mid-rank convention avoids mapping to exactly 0 or 1.

Each sample is then transformed by applying the inverse CDF of the standard normal:

$$\tilde{r}_i^k \;=\; \text{clip}\Big[\, F_{\mathcal{N}}^{-1}\Big(F_{r^k}(r_i^k)\Big),\; -3,\; 3\,\Big], \qquad i = 1, \ldots, B, \tag{11}$$

where $F_{\mathcal{N}}^{-1}$ is the standard normal quantile function and the clipping to $[-3, 3]$ prevents extreme outliers. This yields the optimally transported samples $\{\tilde{r}_i^k\}_{i=1}^B$ that follow the target Gaussian distribution.

4.3 3D-Balanced-R1 Training

We generate full denoising trajectories without gradients for video-level reward evaluation, retaining only two transitions for policy optimization to reduce computation and memory. Let $T = 50$ be the total denoising steps and $H = 8$ the initial camera-conditioned noise-injection steps ($1 \le H < T$). For each rollout, we sample chronological indices

$$s_i^{\text{early}} \sim \mathcal{U}\{0, \ldots, H-1\}, \qquad s_i^{\text{late}} \sim \mathcal{U}\{H, \ldots, T-1\}. \tag{12}$$

We optimize $\mathcal{T}_i = \{T - s_i^{\text{early}}, T - s_i^{\text{late}}\}$, where $t = T - s$ maps generation order to the transition $\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i$ in Sec. 3. The early and late transitions capture camera-noise injection and its subsequent propagation, respectively; both require well-defined stochastic policy likelihoods.

After visibility reweighting (Eq. equation 10) and component-wise OT normalization (Eq. equation 11), we compute group-relative advantages. For prompt c_i of rollout i, let $\mathcal{B}_c$ index a complete group of G rollouts sharing prompt c, as in Sec. 3. Then

$$A^{i,k} = \frac{\tilde{r}_i^k - \mu_{c_i}^k}{\sigma_{\text{rew},k}}, \qquad \mu_c^k = \frac{1}{|\mathcal{B}_c|} \sum_{j\in\mathcal{B}_c} \tilde{r}_j^k \tag{13}$$

Here $\sigma_{\mathrm{rew},k} > 0$ is the standard deviation of $\tilde{r}_i^k$ over the gathered batch. Unlike Eq. equation 4, we center each component within its prompt group but scale it across the batch. Both selected transitions share the combined video-level advantage

$$A^i = \sum_{k \in \mathcal{K}} \lambda_k A^{i,k}, \qquad \lambda_{\mathrm{v}} = 1, \quad \lambda_{\mathrm{r}} = \lambda_{\mathrm{t}} = 0.5. \tag{14}$$

Restricting the clipped surrogate in Eq. equation 6 to $\mathcal{T}_i$ gives

$$\mathcal{L}_{\mathrm{GRPO}} = -\frac{1}{2B} \sum_{i=1}^{B} \sum_{t \in \mathcal{T}_i} \omega_t \min\!\big(\rho_t^i(\theta) A^i,\ \mathrm{clip}(\rho_t^i(\theta), 1-\varepsilon, 1+\varepsilon) A^i\big)\,. \tag{15}$$

Here $\rho_t^i(\theta)$ is the policy ratio in Sec. 3, conditioned on c_i, and $\omega_t \geq 0$ is a fixed additional timestep weight ($\omega_t = 1$ recovers the unweighted surrogate). Retaining the same frozen-reference KL regularization, the full loss is

$$\mathcal{L} = \mathcal{L}_{\mathrm{GRPO}} + \frac{\beta}{2B} \sum_{i=1}^{B} \sum_{t \in \mathcal{T}_i} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid \boldsymbol{x}_t^i, c_i) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid \boldsymbol{x}_t^i, c_i)\big)\,. \tag{16}$$

5 Experiments

5.1 Experimental Setup

Data and evaluation metrics. We follow the World-R1 training and evaluation protocol, using the same training prompts and a held-out test set of 300 prompts. Target camera trajectories are derived from the camera motion specified in each prompt. For evaluation, we generate 81-frame videos at 832×480 resolution with 50 denoising steps. We evaluate 3D-aware generation using a reconstruction-consistency score (Recon) and a camera-trajectory alignment score (Traj), both computed from geometric predictions of Depth Anything 3. Recon is the reconstruction score after visibility-based reweighting in Eq. equation 10, while Traj measures how well the generated camera motion matches the target trajectory. Following World-R1, we use $\mathrm{W3D} = \mathrm{Recon} + \mathrm{Traj}$ as the primary 3D metric and report both components separately. We additionally report trajectory alignment estimated by VGGT-Ω (Traj_Ω) as an alternative measure of camera motion. Video quality is evaluated with VideoHPSv3 (HPSv3) Ma et al. (2025b), and we further report VBench Huang et al. (2024) to assess generalization beyond the reward metrics used during training.

Training settings. For Wan2.1-1.3B, we follow the World-R1 training data and optimization settings, except that the KL coefficient is increased from 0.004 to 0.04 to reduce HPSv3 degradation. Training uses $8\times$H100 GPUs with an inference batch size of 2, group size of 8, and gradient accumulation over 24 steps, corresponding to 48 prompts and 384 generated videos per epoch. Our full 56-epoch run takes approximately 100 hours (800 H100 GPU-hours). For Wan2.2-5B, we use $8\times$H200 GPUs and reduce the inference batch size to 1 while keeping the remaining settings unchanged, resulting in 24 prompts and 192 videos per epoch.

5.2 Comparisons and Analysis

This section compares our method with pretrained video diffusion models, including LTX-2.5-22B, and with World-R1 under matched-epoch and full-training settings. We also evaluate whether the 3D-awareness improvements generalize to standard video-quality metrics. Additional video comparisons are provided in `supplemental.zip`

Comparison with World-R1 and SOTA video diffusion models. Table 1 compares 3D-Balanced-R1 with the pretrained Wan backbones, World-R1, and LTX-2.5-22B. To compare 3D-Balanced-R1 and World-R1 under the same training-data exposure, we adopt an equal-epoch setting of 20 epochs. Under this setting, 3D-Balanced-R1 achieves a W3D score of 1.353, corresponding to a 9.6% gain over the pretrained Wan2.1-1.3B backbone, compared with 6.3% for World-R1, i.e., a $1.5\times$ larger improvement. With full training for 56 epochs, W3D further increases to 1.412, yielding a 14.3% gain over the pretrained backbone and a $2.3\times$ larger improvement than World-R1, while surpassing

Table 1: Generation performance on 300 test prompts from the World-R1 dataset. Ours (equal-epoch) denotes the setting where 3D-Balanced-R1 and World-R1 are trained for the same number of epochs. Parentheses show relative changes over the pretrained model with the same backbone for all trained results. Bold indicates the highest value in each metric.

Method	Backbone	W3D ↑	Recon ↑	Traj ↑	Traj_Ω ↑	HPSv3 ↑
Base	Wan2.1-1.3B	1.235	0.826	0.409	0.412	9.389
Base	Wan2.2-5B	1.205	0.790	0.415	0.417	9.401
Base	LTX-2.5-22B	1.312	0.897	0.414	0.416	9.439
World-R1	Wan2.1-1.3B	1.313 (+6.3%)	0.886 (+7.3%)	0.426 (+4.3%)	0.427 (+3.7%)	10.164 (+8.3%)
World-R1	Wan2.2-5B	1.240 (+2.9%)	0.815 (+3.2%)	0.425 (+2.2%)	0.425 (+2.0%)	9.986 (+6.2%)
Ours (equal-epoch)	Wan2.1-1.3B	1.353 (+9.6%)	0.880 (+6.6%)	0.472 (+15.5%)	0.475 (+15.1%)	9.966 (+6.1%)
Ours (equal-epoch)	Wan2.2-5B	1.326 (+10.0%)	0.865 (+9.6%)	0.460 (+10.8%)	0.466 (+11.8%)	9.725 (+3.4%)
Ours	Wan2.1-1.3B	**1.412** (+14.3%)	**0.932** (+12.9%)	**0.480** (+17.3%)	0.484 (+17.4%)	**10.229** (+8.9%)
Ours	Wan2.2-5B	1.365 (+13.3%)	0.890 (+12.7%)	0.476 (+14.5%)	**0.486** (+16.6%)	10.019 (+6.6%)

Table 2: Training efficiency comparison between Ours and World-R1 for one training epoch (384 rollouts) on 8 GPUs.

Phase	World-R1 (min)	Ours (min)	Speedup
Rollout generation	116.22	92.97	$1.25\times$
Policy update	269.61	16.30	$16.54\times$
Total time	385.83	109.27	$3.53\times$

Table 3: Ablation study of our proposed modules on the World-R1 test set using the Wan2.1-1.3B model.

Variant	Recon ↑	Traj ↑	Traj_Ω ↑	HPSv3 ↑	Total ↑
Ours	**0.932**	**0.480**	**0.484**	10.229	**2.435**
No Vis.	0.828	0.431	0.441	10.202	2.279
No OT	0.845	0.473	0.483	**10.390**	2.357
No Early-Late	0.891	0.437	0.444	9.943	2.322

World-R1 across all reported metrics. By reweighting the reconstruction reward with visibility, 3D-Balanced-R1 also substantially improves camera-trajectory following, with Traj increasing by 17.3% over the pretrained backbone, compared with 4.3% for World-R1, corresponding to a $4.0\times$ larger improvement. The same trend holds when scaling to Wan2.2-5B. World-R1 improves W3D from 1.205 for the pretrained backbone to 1.240, while 3D-Balanced-R1 reaches 1.326 under the equal-epoch setting and 1.365 after full training.

Training efficiency comparison. As shown in Table 2, the efficiency gain mainly comes from our two-step policy update. Both methods generate full 50-step denoising trajectories, but World-R1 optimizes 24 transitions per rollout whereas we retain only two. This reduces memory use, allowing the generation batch size to increase from 1 to 2 and yielding a $1.25\times$ rollout speedup. Optimizing $12\times$ fewer transitions also makes policy updates $16.54\times$ faster, resulting in an overall $3.53\times$ training speedup.

Generalization Evaluation on VBench. We further evaluate 3D-Balanced-R1 on VBench to assess generalization beyond the training objectives. Compared with both the pretrained Wan2.1-1.3B model and World-R1, 3D-Balanced-R1 improves Aesthetic Quality, Imaging Quality, and especially Dynamic Degree, while maintaining comparable Motion Smoothness. Background and Subject Consistency decrease slightly. This trade-off is expected because larger viewpoint and scene changes increase video dynamics while making temporal and background consistency more challenging. Overall, 3D-Balanced-R1 achieves the highest average score of 0.811 across the six reported VBench metrics.

5.3 Ablation Study

Table 3 evaluates the three design choices introduced in our method: visibility-aware rewards, reward normalization with optimal transport (OT), and the two-step policy update. Here, $\text{Total} = \text{W3D} + 0.1\,\text{HPSv3}$. The results demonstrate how each component addresses key challenges in applying GRPO to improve the 3D awareness of video diffusion models.

Visibility reward. Without visibility, camera-trajectory alignment drops substantially, with Traj decreasing from 0.480 to 0.431 and Traj_Ω from 0.484 to 0.441; HPSv3 also decreases from 10.229 to 10.202. This reflects the conflict between reconstruction rewards, which can favor videos with

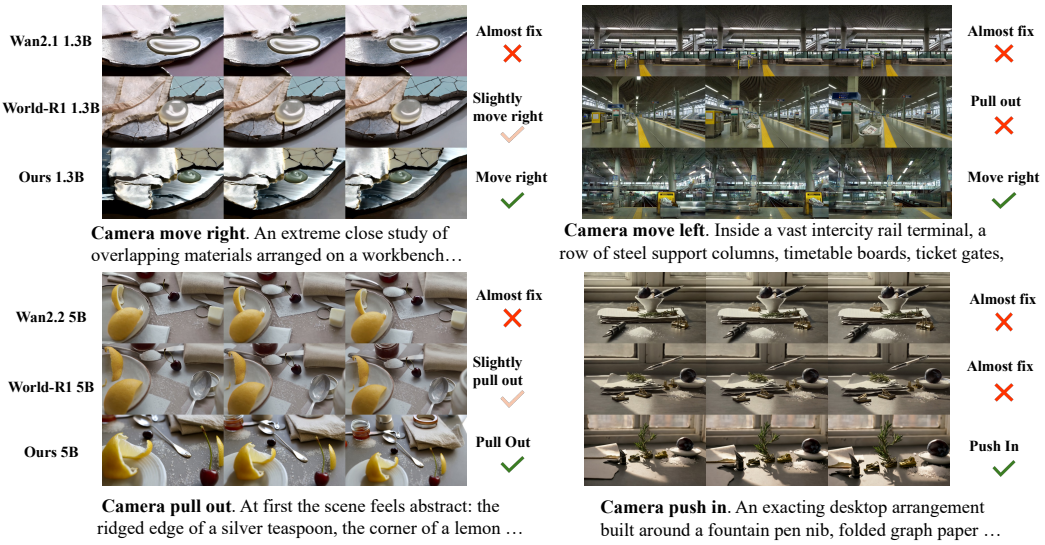


Figure 4: **Visual comparison.** 3D-Balanced-R1 improves camera-motion control on both Wan2.1-1.3B and Wan2.2-5B, better following camera-motion prompts and producing larger viewpoint changes than World-R1 while preserving 3D consistency. This demonstrates that visibility-aware reconstruction rewards effectively mitigate the conflict between 3D reconstruction consistency and camera trajectory. Additional video comparisons are provided in the supplementary HTML included in `supplemental.zip`.

Table 4: VBench evaluation results.

Method	Aesthetic ↑	Image quality ↑	Motion Smooth. ↑	Subject Cons. ↑	BG Cons. ↑	Dynamic ↑	Avg. ↑
Wan2.1-1.3B	0.579	0.636	0.973	0.936	0.955	0.689	0.795
World-R1	0.597	0.649	0.975	**0.952**	**0.964**	0.650	0.798
Ours	**0.598**	**0.668**	**0.976**	0.940	0.937	**0.747**	**0.811**

little camera motion, and trajectory rewards, which encourage useful viewpoint changes. Adding visibility resolves this conflict, and Recon, both trajectory scores, and HPSv3 all improve. These across-the-board gains demonstrate that visibility is an effective way to mitigate the conflicting incentives and improve 3D-aware generation without sacrificing video quality.

Optimal-transport reward normalization. Adding OT raises Total from 2.357 to 2.435, despite a small decrease in HPSv3 from 10.390 to 10.229. This higher overall score is the key result and supports the intended role of OT. By aligning reward distributions with different scales and outlier behavior, OT reduces the dominance of any single reward and improves their joint optimization.

Two-step policy update. Unlike Flash-GRPO's He et al. (2026) single-step update, ours uses an early step for text-guided camera control and a late step to preserve video quality and fine-grained 3D consistency. Compared with the single-step "No Early-Late" variant, our two-step update improves Traj (0.437 to 0.480), Traj_Ω (0.444 to 0.484), HPSv3 (9.943 to 10.229), and Total (2.322 to 2.435).

6 Conclusion and Future Work

We presented 3D-Balanced-R1, an efficient RL framework for improving the 3D awareness of video diffusion models without modifying the model architecture. Visibility-aware reconstruction reweighting mitigates the conflict between reconstruction and camera motion, optimal transport balances heterogeneous reward distributions, and two-step policy updates substantially reduce optimization cost. Together, these designs improve 3D-aware generation performance while preserving general video quality and reducing training overhead across different model scales. With policy optimization largely accelerated, rollout generation becomes the main remaining bottleneck. Combining 3D-Balanced-R1 with few-step video diffusion models is therefore a promising direction for further reducing rollout cost and improving the efficiency of RL post-training.

AI Use Statement. AI-assisted tools were used only for language editing, including correcting grammatical and formatting errors and polishing wording for clarity and readability.

Ethics Statement. This work focuses on improving the training efficiency and 3D awareness of video diffusion models. We do not introduce new datasets involving private or personally identifiable information, nor do we conduct experiments involving human subjects.

Reproducibility Statement. Our implementation is built upon the publicly available training codebases of Flash-GRPO and World-R1. We describe the proposed visibility-aware reward, optimal-transport reward alignment, two-step policy update, and training settings in the main paper. We will release our code and configuration files to enable the community to reproduce our results and apply the proposed method to other video diffusion models.

REFERENCES

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.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Yang Bai, Kaiyuan Liu, Ziyuan Zhuang, Jiahong Zhou, Rongxiang Weng, Xin Chen, Jingang Wang, and Xunliang Cai. Multi-objective and mixed-reward reinforcement learning via reward-decorrelated policy optimization. *arXiv preprint arXiv:2605.13641*, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

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*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575. IEEE, 2023b.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13–23. IEEE, 2025.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in neural information processing systems*, 36:79858–79885, 2023.

Ruiqi Gao*, Aleksander Holynski*, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole*. Cat3d: Create anything in 3d with multi-view diffusion models. *Advances in Neural Information Processing Systems*, 2024.

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.

Xiaoxuan He, Siming Fu, Zeyue Xue, Weijie Wang, Ruizhe He, Yuming Li, Dacheng Yin, Shuai Dong, Haoyang Huang, Hongfa Wang, et al. Flash-grpo: Efficient alignment for video diffusion via one-step policy optimization. *arXiv preprint arXiv:2605.15980*, 2026.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

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.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yirunib8l8.

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. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026b.

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025a.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

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.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024a.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y. Chen, Zhiyuan He, Yuqing Yang, and Bohan Zhuang. World-r1: Reinforcing 3d constraints for text-to-video generation. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=lh0I4rB0I5.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

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. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026a.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26024–26035. IEEE, 2025.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026b.

Xunzhi Xiang, Zixuan Duan, Yabo Chen, Zhengxuan Wei, Guiyu Zhang, Zixiao Gu, Zhe Gao, Haibin Huang, Chi Zhang, Qi Fan, et al. Videoweave: Unlocking geometric consistency in video generation via joint geometry-video modeling. *arXiv preprint arXiv:2606.14162*, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6463–6474. IEEE, 2024.

Canyu Zhao, Hao Chen, Yunze Tong, Yu Qiao, Jiacheng Li, and Chunhua Shen. Marble: Multi-aspect reward balance for diffusion rl. *arXiv preprint arXiv:2605.06507*, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026.

A Additional Qualitative Results

This supplement provides additional qualitative results. Figures 5 and 6 show representative frames from videos generated by 3D-Balanced-R1 with Wan2.1-1.3B on VBench prompts. After fine-tuning, the model still produces dynamic videos while keeping the backgrounds visually consistent across frames. Together with the 3D-awareness gains in Table 1 and the VBench results in Table 4, these examples show that our method improves 3D awareness without sacrificing general video-generation capabilities.

Figures 7–9 provide further qualitative comparisons on the World-R1 test set.

Video comparisons are also provided in `supplement.zip`. After extracting the archive, open the included HTML page to view the videos.



Figure 5: 3D-Balanced-R1 with Wan2.1-1.3B on VBench prompts (part 1).



A drone flying over a snowy forest



A fat rabbit wearing a purple robe walking through a fantasy landscape



A motorcycle accelerating to gain speed



A motorcycle turning a corner



A shark is swimming in the ocean

Figure 6: 3D-Balanced-R1 with Wan2.1-1.3B on VBench prompts (part 2).



Camera move left. A vast desert canyon opens in stepped layers …



Camera push in. A high alpine basin just after a summer storm…

Figure 7: Wan2.1-1.3B results on the World-R1 test set (part 1).

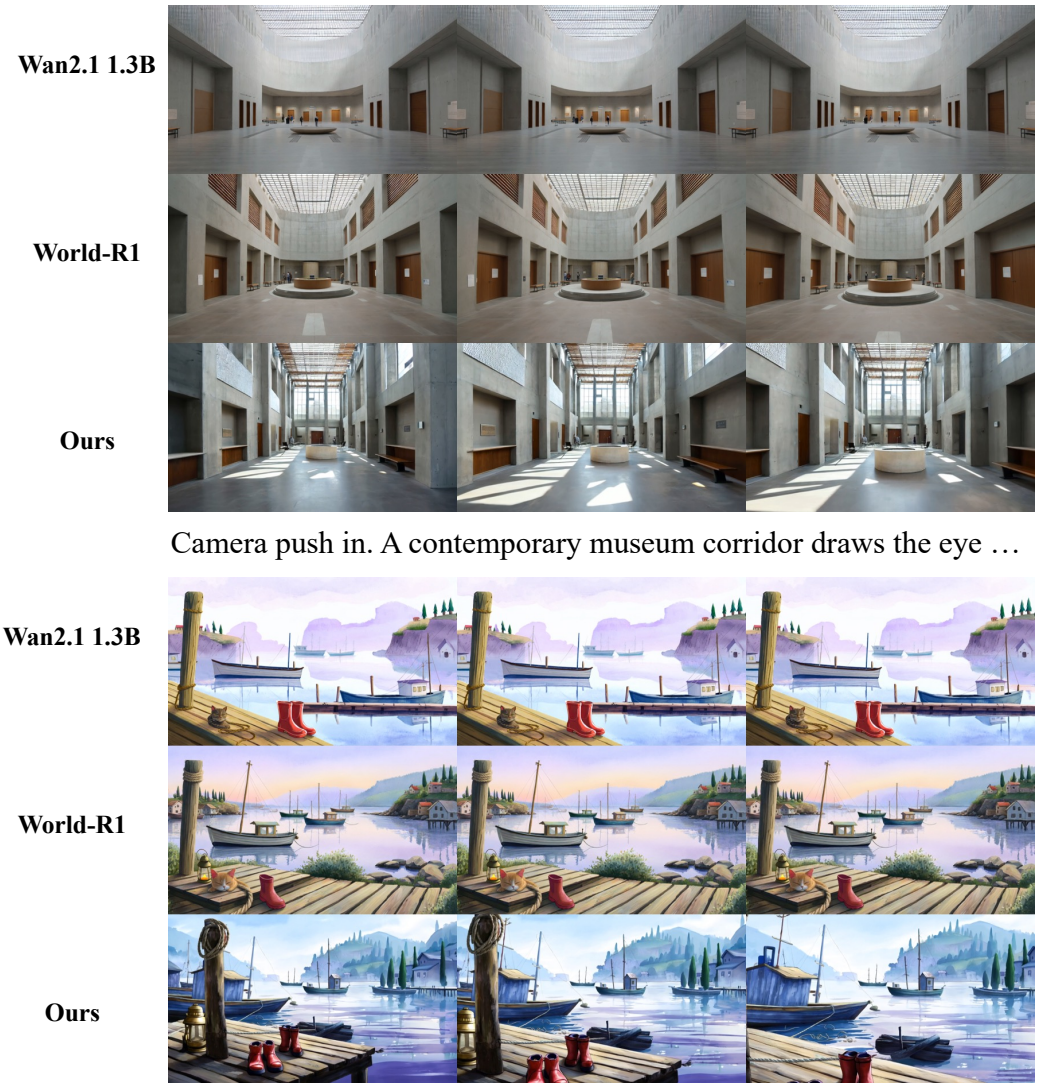
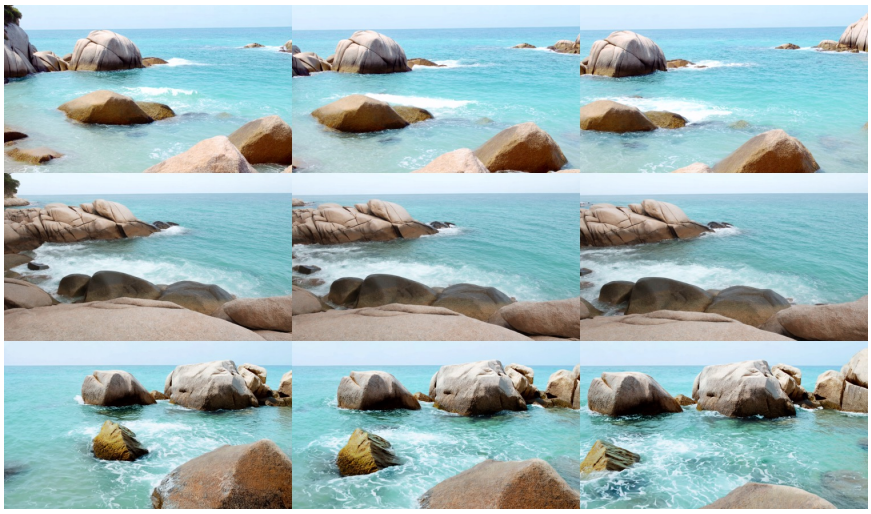


Figure 8: Wan2.1-1.3B results on the World-R1 test set (part 2).



Camera pan right. A serene coastline unfolds…



Camera push in. A dense downtown avenue at early dusk stretches between tall office towers

Figure 9: Wan2.2-5B results on the World-R1 test set.