

IC-World: In-Context Generation for Shared World Modeling

Fan Wu¹, Jiacheng Wei¹, Ruibo Li¹, Yi Xu³, Junyou Li²,
Deheng Ye^{2(✉)}, and Guosheng Lin^{1(✉)}

¹ Nanyang Technological University, Singapore
fan011@e.ntu.edu.sg, gslin@ntu.edu.sg

² Tencent, China

³ Goertek Alpha Labs, China

<https://github.com/wufan-cse/IC-World>

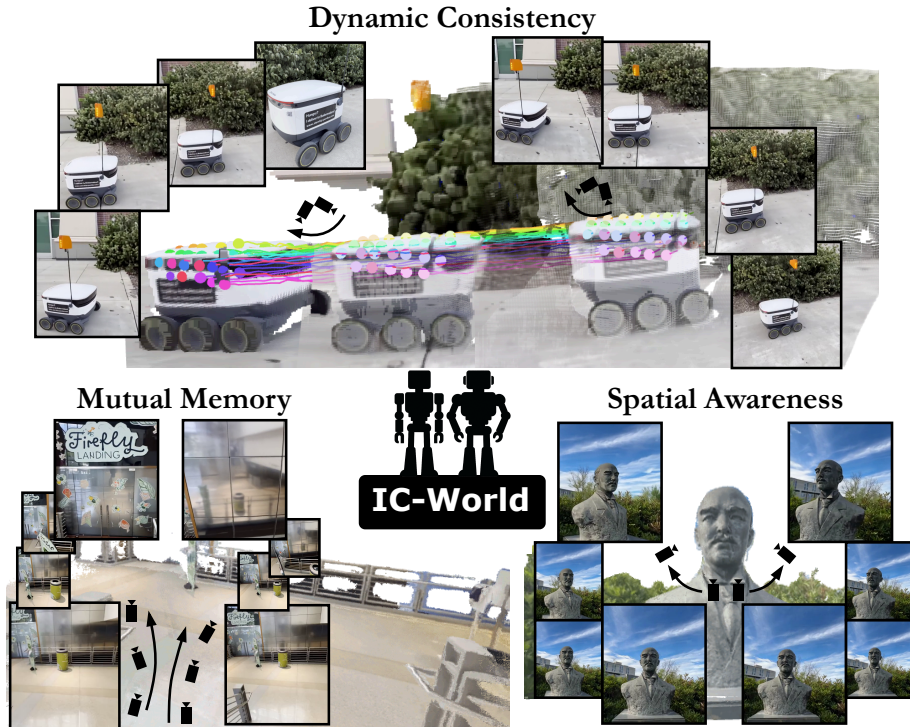


Fig. 1: IC-World is a shared world modeling framework, which aims at generating N shared world videos from N input images, respectively, where all input images are snapshots of the same underlying world. Here we present cases of two generated videos with two unconstrained camera trajectories. This figure illustrates three core capabilities of IC-World: (1) **Dynamic consistency** - object motions remain coherent across different views. (2) **Mutual memory** - scenes reappear consistently, as shown by the complete words above the gate generated in the left view. (3) **Spatial awareness** - both foreground objects and background scenarios exhibit strong geometry consistency.

Abstract. Video-based world models have recently garnered increasing attention for their ability to synthesize diverse and dynamic visual environments. In this paper, we focus on shared world modeling, where a model generates multiple videos from a set of input images, each representing the same underlying world in different camera poses. We propose IC-World, a novel generation framework, enabling parallel generation for all shared world input images via activating the inherent in-context generation capability of large video models. We further finetune IC-World via reinforcement learning, Group Relative Policy Optimization, together with two proposed novel reward models to enforce scene-level geometry consistency and object-level motion consistency among the set of generated videos. Extensive experiments demonstrate that IC-World substantially outperforms state-of-the-art methods in both geometry and motion consistency. To the best of our knowledge, this is the first work to systematically explore the shared world modeling problem with video-based world models.

Keywords: Video Generation · In-context learning · Reinforcement Learning

1 Introduction

With the recent advances in video-based world models [7, 15, 19, 23, 34, 42, 43], shared world modeling has emerged as an important extension. We define shared world modeling as the task of generating a set of videos that are both spatially and temporally consistent, where each video corresponds to a distinct input image of the same underlying world and unlimited camera trajectories, as shown in Figure 1. This problem forms a foundation for numerous critical applications, including multiplayer video gaming [5] and multi-robot coordination simulation, all of which involve multiple independent cameras that must coherently represent a single shared world.

Existing video-based world models contrast with this task, as they typically generate videos from a single viewpoint, i.e., image-to-video (I2V). Directly extending them to this task remains non-trivial, since sequentially and independently generating videos of the shared world input images would introduce complex dependencies across generated videos to ensure spatial and temporal consistency. Moreover, hallucination effects [8, 33] observed in these models further amplify inconsistencies between generated videos. Another line of works focus on domain-specific world modeling. Multi-view video generation for autonomous driving [32, 49, 50] aims to generate consistent cross-camera views to simulate the views of ego-vehicles. 4D video generation [35] aims to generate consistent object-centric videos with continuous viewpoints. However, these works either use a preset multi-camera system or focus on object-centric generation, limiting their usage in advanced world modeling applications, such as multiplayer video gaming, which requires scene-level generation with multiple free cameras.

To address shared world modeling, we introduce IC-World, a novel generation framework that leverages the inherent in-context generation capabil-

ity [12] of large video models to generate shared world videos. As shown in Figure 1, our method effectively performs shared world modeling, exhibiting mutual memory across generated views. To further enhance consistency, we fine-tune IC-World with reinforcement learning, Group Relative Policy Optimization (GRPO), and equip it with two novel reward models designed to ensure scene-level geometry consistency and object-level motion consistency across generated videos. Extensive experiments demonstrate that the proposed framework significantly surpasses state-of-the-art video-based world modeling methods on our benchmark and exhibit strong visual quality as shown in Figure 1.

In summary, our contributions are as follows:

1. To the best of our knowledge, this is the first work to formulate shared world modeling as a new setting for video-based world modeling and to establish a comprehensive benchmark with well-designed evaluation metrics.
2. We proposed IC-World, a novel instructional generation framework to activate the inherent in-context generation capability of large video models to generate multiple shared-world videos in a parallel manner.
3. We introduce GRPO-based reinforcement finetuning with two novel reward models to further enforce scene-level geometry and object-level motion consistency.

2 Related Works

Video-based world models. The emergence of Sora has further highlighted the potential of large video generation models as world simulators. Recent advancements have primarily focused on text-to-video (T2V) [6, 7, 25], and image-to-video (I2V) [1, 15, 19, 34, 43, 45]. However, conventional I2V methods struggle to address the shared world modeling, as they fail to guarantee that all generated videos depict the same underlying world. Other extension tasks of video-based world modeling, including multi-view video generation [24, 32, 39, 49, 50], and 4D video generation [35, 38, 41], typically rely on training additional modules to enforce cross-view consistency. In contrast, IC-World achieves this by leveraging the model’s inherent in-context reasoning, without introducing any extra modules.

Instructional in-context generation. Recent studies in image generation [10, 21, 31, 48] have revealed that large-scale diffusion transformers (DiTs) inherently possess in-context generation capabilities, which can be activated through appropriately designed instructional prompts. IC-LoRA [21] demonstrates that existing text-to-image DiTs are capable of performing in-context generation without additional fine-tuning. Similarly, IC-Edit [48] introduces an instruction-based framework that exploits the latent comprehension and generation abilities of large-scale DiTs for image editing. Beyond text-to-image, Fei et al. [12] extend this to video models, demonstrating the remarkable adaptability of the idea. In this paper, we extend this inherent in-context capability to shared world modeling and propose a novel framework to generate consistent videos.

Reinforcement finetuning for video models. Reinforcement learning has recently emerged as an effective paradigm for optimizing large-scale generative models [29, 44, 46, 47], particularly for enhancing reasoning and decision-making capabilities. DeepSeek-R1 [14] introduces Group Relative Policy Optimization (GRPO), which leverages verifiable reward signals to estimate relative advantages among model responses, thereby substantially improving reasoning performance. Following this advancement, GRPO-based fine-tuning has been successfully extended to a variety of multimodal domains, including video understanding [13, 36], multi-image grounding [4], and visual generation [11, 44]. These developments collectively demonstrate the versatility of GRPO in optimizing the complex visual generation task. Building upon this, we design two novel reward models to enforce scene-level geometry consistency and object-level motion consistency across the generated videos for shared world modeling.

3 Methodology

Problem setup. Shared world modeling aims to output a set of $N > 1$ videos $\mathcal{V} = \{v_i\}_{i=1}^N$ conditioned on a corresponding set of N input images $\mathcal{I}$ and a shared text prompt c without a constrained camera trajectory for each input image. The input images capture the same world at the same time from different camera views; in other words, the input images are multi-view and fixed-time snapshots of an underlying world described by c , including its static scenes and dynamic objects. In this study, we focus on the simplest but non-trivial case with $N = 2$ since we observe that even this setting remains challenging for existing methods demonstrated in our experiments.

3.1 Preliminary

DanceGRPO [44] is a reinforcement learning framework that extends GRPO to visual generation tasks. The central idea is to reinterpret the denoising process as a Markov Decision Process (MDP), where each timestep represents a state transition and the reward of the final output. Formally, for a given prompt c , the state at denoising timestep t is defined as $s_t = (c, t, z_t)$, where z_t denotes the latent at t . The policy model π_θ , which is the video model with parameters θ , will generate $M > 1$ outputs $\mathcal{O} = \{o_i\}_{i=1}^M$ and governs the transition from z_t toward predicting lower-noise latent z_{t-1} with the probability policy function:

$$\pi_\theta(z_{t-1,i} \mid s_{t,i}) := p(z_{t-1,i} \mid z_{t,i}, c), i \in [1, M], \quad (1)$$

where each output o_i contains $\{v_i\}_{i=1}^N$ videos of N input images. A learned or rule-based reward model will then compute reward $\{r_i\}_{i=1}^M$ for each output. Then the training of GRPO maximizes the following objective $L(\theta)$:

$$L(\theta) = \mathbb{E} [\min(\rho_{t,i} A_i, \text{clip}(\rho_{t,i}, 1 - \epsilon, 1 + \epsilon) A_i)], \quad (2)$$

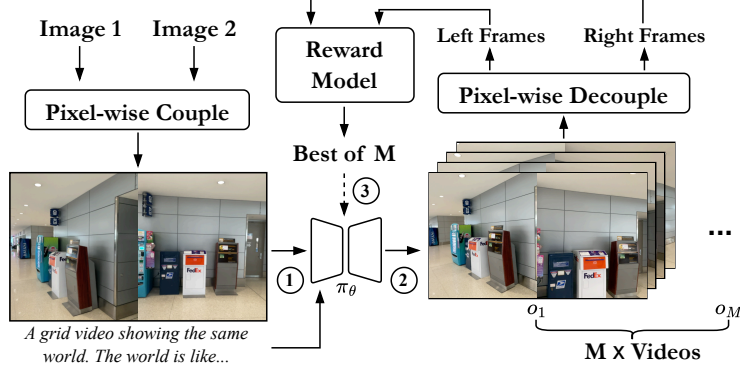


Fig. 2: Overview of IC-World. Here we present the case of $N = 2$. (1) The input images are first concatenated pixel-wise, accompanied by an in-context prompt, to perform I2V generation. (2) The model π_θ then generates M candidate videos conditioned on the same input. (3) Each candidate video is decoupled pixel-wise into sub-videos for reward calculation. Then we find the best candidate to calculate gradients for fine-tuning.

where $\rho_{t,i} = \frac{\pi_\theta(z_{t-1,i}|s_{t,i})}{\pi_{\theta_{\text{prev}}}(z_{t-1,i}|s_{t,i})}$ is the policy ratio between the previous sample and the current sample, and the group-normalized advantage is computed as:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^M)}{\text{std}(\{r_j\}_{j=1}^M)}. \quad (3)$$

This normalization allows each sample’s contribution to be evaluated within its group, enhancing training stability.

3.2 In-Context Generation

Shared world modeling aims to generate N videos given a shared text prompt c and N input images. Conventional I2V models generate each video sequentially and independently, making it difficult to ensure the dependencies required by output videos, where all videos share the same world. This inconsistency is further amplified by the inherent hallucination tendencies of I2V models. To address this limitation, we introduce a minor yet effective modification to conventional video diffusion transformers. Specifically, we design an in-context prompt template for shared world modeling as follows: “A *grid video showing the same, synchronized world captured from different camera poses. The world is like: <WORLD>.*” $\langle \text{WORLD} \rangle$ denotes a placeholder that is substituted with the textual description of the underlying world. Alongside the prompt, we down-scale each input image and combine them pixel-wise into a single grid image as shown in Figure 2, in a way to serve as the model input and perform parallel generation for all input images. After inference, we decouple the generated large grid videos into N small sub-videos pixel-wise. With this generation framework, the inherent capability of large video models as the world’s in-context generator can be activated to output consistent videos.



(a) **Geometry consistency** under the static scene and dynamic camera. (b) **Motion consistency** under the dynamic scene and static camera.

Fig. 3: Reward visualization. Our proposed reward models are able to choose the best candidate that aligns well with human visual perception. Better view in zoom-in mode.

3.3 Shared World Reward Models

To further enhance consistency among generated sub-videos, we proposed two reward models for consistency on scene-level geometry and object-level motion. For clarity, we illustrate the case of $N = 2$.

Geometry consistency reward model. To quantify geometric consistency between two generated sub-videos, we propose a point cloud-based [28, 37] 3D reconstruction reward model. Each video is preprocessed through the Pi3 model to obtain dense 3D point clouds $\mathcal{P}_1 \in \mathbb{R}^{N_1 \times 3}$ and $\mathcal{P}_2 \in \mathbb{R}^{N_2 \times 3}$, respectively. Since the two videos are captured from different camera perspectives, their reconstructions reside in distinct coordinate systems. To align them, we register $\mathcal{P}_1$ to $\mathcal{P}_2$ using Lepard [26], a learned point cloud registration method that robustly handles partial overlaps and noisy reconstructions through keypoint-based feature matching. After alignment, geometric similarity is quantified by the symmetric Chamfer distance:

$$D_g(\mathcal{P}_1, \mathcal{P}_2) = \frac{1}{2} \left[\frac{1}{N_1} \sum_{x \in \mathcal{P}_1} \min_{y \in \mathcal{P}_2} \|x - y\|_2 + \frac{1}{N_2} \sum_{y \in \mathcal{P}_2} \min_{x \in \mathcal{P}_1} \|y - x\|_2 \right], \quad (4)$$

$$r_g = \exp(-D_g) \in (0, 1],$$

where r_g is the geometry consistency reward. As shown in Figure 3a, lower D_g yields higher reward r_g , encouraging the model to generate videos with consistent 3D geometry.

Motion consistency reward model. To evaluate motion consistency between two videos, we employ a 3D point tracking-based reward model built upon SpatialTracker [40]. Given two videos, we uniformly sample T_p frames for each to predict camera extrinsics $\mathcal{C}_1, \mathcal{C}_2$ and 3D point trajectories $\mathcal{B}_1 = \{b_{1,i}(t_p) \in \mathbb{R}^3\}$ and $\mathcal{B}_2 = \{b_{2,i}(t_p) \in \mathbb{R}^3\}$, $i \in [1, B]$, $t_p \in [1, T_p]$, where each trajectory is represented in its own camera coordinate system. We use the same definition of camera system as CameraCtrl [16] and align coordinate systems of video pair according to camera extrinsics: $\mathcal{C}_a = \mathcal{C}_1 \cdot \mathcal{C}_2^{-1}$, then align the tracks:

$$\hat{b}_{1,i}(t_p) = \mathcal{C}_{a,rotate} \cdot b_{1,i}(t_p) + \mathcal{C}_{a,trans}, \quad (5)$$

where $\mathcal{C}_{a,rotate}$ and $\mathcal{C}_{a,trans} \in \mathbb{R}^3$ are the rotation and translation components of $\mathcal{C}_a$. We establish point correspondences by computing temporal-average positions

Algorithm 1 IC-World: In-Context Generation for Shared World Modeling

```

1: Input: Initial policy video model  $\pi_\theta$ ; reward scales  $\lambda_g, \lambda_m$ ; training dataset  $\mathcal{D}_{\text{train}}$ ;
   training timestep ratio  $\tau$ ; denoising steps  $T$ ; generation group size  $M$ ; number of
   player  $N$ ; learning rate  $\eta$ .
2: Output: Video model with optimized parameters  $\theta^*$ .
3: for each training step do
4:   Subsample training data  $\mathcal{D}_{\text{sub}}$  from  $\mathcal{D}_{\text{train}}$  ;
5:   Update old policy:  $\pi_{\theta_{\text{prev}}} \leftarrow \pi_\theta$  ;
6:   for each  $(c, \mathcal{I}) \in \mathcal{D}_{\text{sub}}$  do
7:     #  $M$  times I2V inference
8:     Pixel-wise couple  $N$  input images  $\mathcal{I}$  ;
9:     Generate  $M$  samples  $\{o_i\}_{i=1}^M \sim \pi_{\theta_{\text{prev}}}(\cdot | c, \mathcal{I})$  ;
10:    Pixel-wise decouple  $o_i$  to  $N$  videos ;
11:    # Reward calculation
12:    Calculate  $\{r_{g,i}\}_{i=1}^M$  with Equation (4) ;
13:    Calculate  $\{r_{m,i}\}_{i=1}^M$  with Equation (7) ;
14:    Rewards  $\{r_i\}_{i=1}^M \leftarrow \{\lambda_g r_{g,i} + \lambda_m r_{m,i}\}_{i=1}^M$  ;
15:    for each sample  $i \in \{1..M\}$  do
16:      Calculate advantage  $A_i$  with Equation (3) ;
17:      # Gradient calculation
18:      Calculate objective  $L(\theta)$  with Equation (2) ;
19:      Subsample  $[\tau T]$  steps  $T_{\text{sub}} \subset \{1..T\}$  ;
20:      for each  $t \in T_{\text{sub}}$  do
21:        Gradient ascent:  $\theta \leftarrow \theta + \eta \nabla_\theta L$  ;
22: Save the optimized parameters  $\theta^*$ .

```

$\bar{b}_i = \frac{1}{T_p} \sum_{t_p=1}^{T_p} b_i(t_p)$ and defining a matching function $\delta : \mathcal{B}_1 \rightarrow \mathcal{B}_2$ such that:

$$\delta(i) = \arg \min_{j \in [1, B]} \|\bar{b}_{1,i} - \bar{b}_{2,j}\|_2. \quad (6)$$

The motion distance is computed as the Euclidean distance:

$$D_m(\mathcal{B}_1, \mathcal{B}_2) = \frac{1}{BT_p} \sum_{i=1}^B \sum_{t_p=1}^{T_p} \|\hat{b}_{1,i}(t_p) - b_{\delta(i)}(t_p)\|_2, \quad (7)$$

$$r_m = \exp(-D_m),$$

where r_m is the motion consistency reward. As shown in Figure 3b, higher rewards encourage the model to generate consistent 3D motion patterns across videos.

Algorithm framework. Our method is outlined in Figure 2 and Algorithm 1. For each text prompt and input image set $(c, \mathcal{I})$, we first combine the N input images into a single large input image, then generate M candidate large grid video samples. For each grid video sample, we split it into N sub-videos, then compute the rewards with the two proposed reward models. The final reward is the weighted average of the results of two reward models. We use it to compute the objective and update the policy model through gradient ascent. This process

Methods	M-FID ↓	CLIP ↑	VLM ↑	Geometry _{0.1} ↑	Geometry _{0.5} ↑	Geometry _{0.7} ↑
CogVideoX-I2V-5B	96.1624	0.6208	<u>0.7684</u>	0.6727	0.7150	<u>0.6909</u>
Wan2.1-14B	103.5042	<u>0.6204</u>	0.7193	0.6898	<u>0.7171</u>	0.6818
EasyAnimate-12B	173.8106	0.6181	0.7559	0.6911	0.4775	0.1963
DynamiCrafter-1024	136.1273	0.6089	0.5609	<u>0.6940</u>	0.2837	0.0654
VideoCrafter-1024	170.9313	0.6153	0.7656	0.6783	0.5930	0.3845
LTX-Video-13B	115.3138	0.6170	0.7121	0.6825	0.7025	0.6393
IC-World	<u>98.8324</u>	0.6170	0.7822	0.6980	0.7217	0.6994

Table 1: Shared world modeling comparison on static scene and dynamic camera. Numbers in bold indicate the best performance, while underlined numbers represent the second-best. The absolute geometry score difference appears relatively small since the inconsistency is mostly in a local region, while the metric averages over the entire generated scene and the visual difference remains significant. Please refer to Appendix for illustrative examples.

Methods	M-FID ↓	CLIP ↑	VLM ↑	Motion ₁₀ ↑	Motion ₂₀ ↑	Motion ₃₀ ↑
CogVideoX-I2V-5B	202.9558	0.6412	0.7915	0.7853	0.7963	0.8018
Wan2.1-14B	<u>185.2354</u>	<u>0.6405</u>	0.7949	<u>0.8240</u>	<u>0.8344</u>	<u>0.8400</u>
EasyAnimate-12B	231.6344	0.6334	0.7719	0.7636	0.7715	0.7743
DynamiCrafter-1024	216.6677	0.6192	0.7392	0.8101	0.8153	0.8167
VideoCrafter-1024	248.3784	0.6279	0.7916	0.7946	0.8036	0.8063
LTX-Video-13B	200.0291	0.6361	0.7792	0.7957	0.8081	0.8121
IC-World	157.6099	0.6400	0.7958	0.8360	0.8466	0.8500

Table 2: Shared world modeling comparison on dynamic scene and static camera.

progressively refines the model parameters θ to obtain an optimized θ^* that produces geometrically and temporally consistent videos.

4 Experiments

4.1 Experimental Setup

Implementation details. We employ Wan2.1-14B [34] as the foundational model and we use the step-distill version [9] to initialize it, which enables us to perform 4-steps inference to generate high-quality video. In our experiments, we consider the simplest but non-trivial setting $N = 2$, where both the size of input images and output videos are 2, as we observe that existing baselines perform poorly in this setting. To evaluate scalability, we additionally report quantitative and qualitative results for $N > 2$ in Appendix 6. We found that LoRA finetuning [20] produces more coherent results and a more stable training process compared with finetuning all parameters; thus, we opt to implement our algorithm based on LoRA with rank of 64. All experiments are conducted on an 8 H20 GPUs (each with 95GB GPU memory) platform, and the finetuning of our method lasts for 40 hours with group size $M = 16$ and training step up to 200. Reward scales λ_g and λ_m are both set to 0.5. We use the AdamW [30]

Models (I2V)	Subject Consist.	Background Consist.	Motion Smooth.	Dynamic Degree	Aesthetic Quality	Imaging Quality	Weighted Average
<i>Commercial Closed Source Models.</i>							
Gen-4	93.23	96.79	98.99	<u>55.20</u>	61.77	70.41	<u>80.89</u>
<i>Open Source Models.</i>							
Wan2.1-14B	94.86	97.07	97.90	51.38	<u>64.75</u>	70.44	80.82
DynamiCrafter-1024	<u>95.69</u>	<u>97.38</u>	97.38	47.40	66.46	69.34	80.50
VideoCrafter-1024	97.86	98.79	98.00	22.60	60.78	71.68	78.84
CogVideoX-5B	94.34	96.42	<u>98.40</u>	33.17	61.87	70.01	78.61
IC-World	94.22	95.54	97.26	72.36	61.05	<u>70.49</u>	81.15

Table 3: VBench metrics results. The weighted average follows the official weighting scheme defined in VBench.

Settings	Foundation model	M-FID ↓	CLIP ↑	VLM ↑	Geometry _{0.1} ↑	Geometry _{0.5} ↑	Geometry _{0.7} ↑
Zero-shot wo/ IC-generation	LTX-Video-13B	115.3138	0.6170	0.7121	0.6825	0.7025	0.6393
Zero-shot w/ IC-generation		94.1303	0.6186	0.7521	0.6845	0.6784	0.6579
Zero-shot wo/ IC-generation	Wan2.1-14B-distill	103.5042	0.6204	0.7193	0.6898	0.7171	0.6818
Zero-shot w/ IC-generation		99.1513	0.6174	0.7818	0.6918	0.7199	0.7006

(a) Results comparison on static scene and dynamic camera.

Settings	Foundation model	M-FID ↓	CLIP ↑	VLM ↑	Motion ₁₀ ↑	Motion ₂₀ ↑	Motion ₃₀ ↑
Zero-shot wo/ IC-generation	LTX-Video-13B	200.0291	0.6361	0.7792	0.7957	0.8081	0.8121
Zero-shot w/ IC-generation		235.7100	0.6328	0.7822	0.8224	0.8265	0.8280
Zero-shot wo/ IC-generation	Wan2.1-14B-distill	185.2354	0.6405	0.7949	0.8240	0.8344	0.8400
Zero-shot w/ IC-generation		160.0052	0.6402	0.7981	0.8342	0.8441	0.8477

(b) Results comparison on dynamic scene and static camera.

Table 4: Ablation study on in-context generation with different foundation models. The results demonstrate the effectiveness of our designs in activating the inherent in-context generation capability of large video models.

optimizer with a learning rate of 1×10^{-5} . Detailed setting of hyper-parameters can be found in Appendix 6.

Datasets. In our experiments, we evaluate all methods under two benchmark settings. The first setting involves static scenes with dynamic cameras, designed to assess scene-level geometric consistency. The second setting uses dynamic scenes with static cameras, focusing on object-level motion consistency. For the first setting, we sample 1K training and 1K testing data from DL3DV-10K [27], a large-scale scene dataset containing 10,510 videos captured from diverse real-world locations. For the second setting, there is currently no large-scale real-world dataset that provides multiple synchronized views of articulated objects under static cameras. Therefore, we adopt MultiCamVideo [2], a multi-camera synchronized video dataset rendered with Unreal Engine 5. It comprises 13.6K dynamic scenes, each recorded from 10 different camera viewpoints. Similarly, we sample 1K training and 1K testing data from this dataset. For all sampled data, we use Qwen2.5-VL-32B [3] to automatically generate world descriptions, covering both static scene attributes and dynamic object behaviors. Some ex-

Settings	Train. data	M-FID ↓	CLIP ↑	VLM ↑	Geometry _{0.1} ↑	Geometry _{0.5} ↑	Geometry _{0.7} ↑
LoRA w/ SFT	1K	101.1831	0.6150	0.7621	0.6933	0.7079	0.6891
LoRA w/ SFT	2K	100.9812	0.6168	0.7682	0.6913	0.7121	0.6901
LoRA w/ SFT+GRPO	1K	<u>99.1012</u>	0.6173	0.7844	0.7002	<u>0.7193</u>	0.7013
LoRA w/ GRPO	1K	98.8324	<u>0.6170</u>	<u>0.7822</u>	<u>0.6980</u>	0.7217	<u>0.6994</u>

(a) Results comparison on static scene and dynamic camera.

Settings	Train. data	M-FID ↓	CLIP ↑	VLM ↑	Motion ₁₀ ↑	Motion ₂₀ ↑	Motion ₃₀ ↑
LoRA w/ SFT	1K	170.0154	<u>0.6400</u>	0.7909	0.8288	0.8381	0.8426
LoRA w/ SFT	2K	166.4103	0.6398	<u>0.7911</u>	0.8300	0.8368	0.8447
LoRA w/ SFT+GRPO	1K	158.3412	0.6404	0.7910	<u>0.8310</u>	0.8484	<u>0.8491</u>
LoRA w/ GRPO	1K	157.6099	<u>0.6400</u>	0.7958	0.8360	<u>0.8466</u>	0.8500

(b) Results comparison on dynamic scene and static camera.

Table 5: Ablation study on training strategy. All settings are based on Wan2.1-14B and the IC-generation framework. For LoRA w/ SFT+GRPO, we first use SFT for warm-up training then utilize GRPO. The results show that SFT requires more data to obtain strong performance compared with GRPO. GRPO alone achieves performance comparable to SFT+GRPO while being more computationally efficient.

amples of input data can be found in Appendix 7, including input images and the full corresponding prompt.

Baselines. As discussed earlier, other lines of video-based world modeling works, such as multi-view video generation and 4D video generation, either rely on fixed camera trajectories or focus on object-centric generation, and their designs are highly task-specific. Consequently, it is unfair to evaluate them under our setting, which is camera trajectory-free and not limited to object-level generation. Therefore, we compare our approach primarily with general video-based world modeling baselines. We establish two categories of baselines: (1) Foundational models, including CogVideoX-5B [45], Wan2.1-14B [34], and LTX-Video-13B [15]; and (2) Enhanced algorithms designed to improve generation quality on foundational models, including VideoCrafter-1024 [6], DynamiCrafter-1024 [42], and EasyAnimate-12B [43]. For all baselines, we use largest available open-source versions.

Evaluation metrics. Evaluating the consistency of the generated shared world videos is challenging without ground truth data. We employ the following metrics:

M-FID score. For a generated set of N videos, we compute the mutual Fréchet Inception Distance [18] (M-FID) to measure their perceptual discrepancy.

CLIP score [17] evaluates the similarity between the single large video, which is a pixel-wise concatenation of N videos, against the text prompt.

VLM score. We use Qwen2.5-VL-32B [3] to ask whether the generated N videos represent the same underlying world, ranking the consistency level from 0 to 10.

Geometry consistency score. We assess geometry consistency across N videos by reconstructing 3D point clouds for each video using Pi3 [37], then retaining only high-confidence points with three different confidence levels (0.1, 0.5, 0.7). We use Lepard [26] to perform point clouds registration for a pair of videos. Af-

Settings	M-FID ↓	CLIP ↑	VLM ↑	Geometry _{0.1} ↑	Geometry _{0.5} ↑	Geometry _{0.7} ↑
w/o Geometry & Motion	101.2243	0.6150	0.7453	0.6900	0.7180	0.6891
w/ Geometry	97.7391	0.6172	<u>0.7810</u>	<u>0.6937</u>	0.7270	0.7052
w/ Motion	99.1874	0.6169	0.7798	0.6898	<u>0.7253</u>	0.6894
w/ Geometry & Motion	<u>98.8324</u>	<u>0.6170</u>	0.7822	0.6980	0.7217	<u>0.6994</u>

(a) Results comparison on static scene and dynamic camera.

Settings	M-FID ↓	CLIP ↑	VLM ↑	Motion ₁₀ ↑	Motion ₂₀ ↑	Motion ₃₀ ↑
w/o Geometry & Motion	165.4712	0.6399	0.7926	0.8280	0.8364	0.8411
w/ Geometry	159.5929	0.6401	<u>0.7962</u>	0.8322	0.8380	0.8412
w/ Motion	<u>159.1588</u>	0.6403	0.7967	<u>0.8347</u>	<u>0.8445</u>	<u>0.8479</u>
w/ Geometry & Motion	157.6099	0.6400	0.7958	0.8360	0.8466	0.8500

(b) Results comparison on dynamic scene and static camera.

Table 6: Ablation study on reward models. All settings are based on Wan2.1-14B and the IC-generation framework. For w/o Geometry & Motion, we use video-text alignment reward model from DanceGRPO [44]. The results empirically demonstrate that the proposed reward models obtain substantial consistency improvement across both settings.

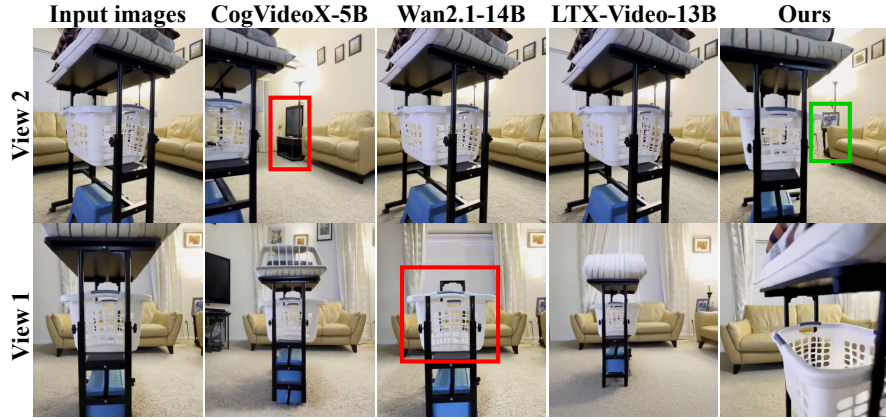
ter registration, we compute the Chamfer distance D_g between registered point cloud pairs and the final score is reversed by $\exp(-D_g)$. If $N > 2$, we consider pair-wise combinations and compute the average over all the combinations as the final score.

Motion consistency score. To compute motion consistency between a pair of videos, we first uniformly sample frames at an interval of 5. We deploy SpatialTrackerV2 [40] to predict the 3D point tracks at three density levels (10, 20, 30, with higher values indicating denser tracking). We align the tracks according to the two videos’ camera extrinsics estimated by SpatialTrackerV2, then we calculate the Euclidean distance D_m between the aligned tracks. The final score is reversed by $\exp(-D_m)$. For the situation $N > 2$, we calculate pairwise scores for all video combinations and report their average.

4.2 Comparison with Existing Methods

Here we present the main comparison results, and we further present a user study in Appendix [6].

Quantitative results. All baselines in Table [1] and Table [2] perform N times sequential generation for N input images. The results show our method achieves the highest geometry and motion consistency across both settings, surpassing all baselines by clear margins. For broader and comprehensive comparison, we implement IC-World on VBench [22] and the results are shown in Table [3]. All the baseline results are reported by official VBench. Table [3] shows that IC-World achieves an optimal weighted average score (81.15), outperforming leading closed source and open source models.



Static scene + Dynamic camera: *A room where a laundry basket fits under an ironing board.*



Dynamic scene + Static camera: *A barista wearing a cap is preparing a cocktail in a kitchen.*

Fig. 4: Visual comparisons. Notably, in the dynamic scene, Wan2.1 fails to maintain spatial alignment, where the man’s hands appear at a higher level in view 1 than in view 2. In contrast, our method can generate consistent results. More comparison results can be found in Appendix 9.

Qualitative results. We visualize our proposed IC-World and baseline methods in Figure 4. In the static scene, our method accurately maintains the geometry consistency across views, with consistent object positioning and perspective (green box). While in the dynamic scene, our method preserves the pouring action of the barista with temporal consistency from different viewpoints. More comparison results can be found in Appendix 9. Some samples of our results on the test set are shown in Figure 6.

4.3 Ablation Studies

Our ablation study systematically validates the contribution of each component in IC-World.

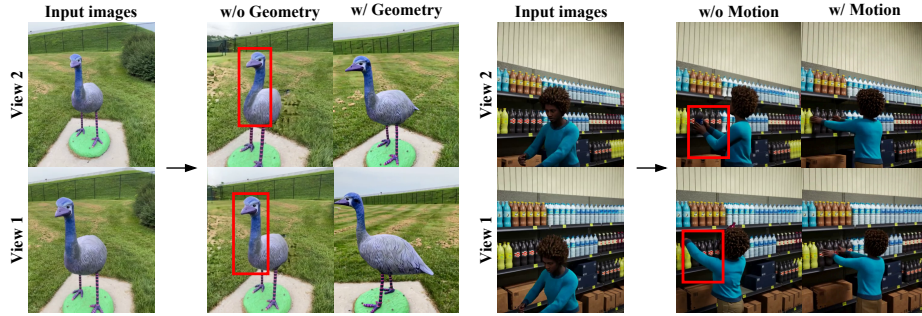


Fig. 5: Ablation on the two reward models. We compare generations without and with the geometry reward (top) and the motion reward (bottom). Introducing the geometry reward enforces geometry consistency across views, while the motion reward improves motion consistency, reducing unnatural motion artifacts and ensuring synchronized movement.



Fig. 6: Generated results visualization on our two test sets. Our method can generate consistent and high-quality results.

In-context generation. Table 4 shows that even in the zero-shot setting, our proposed in-context framework is effective in activating the inherent capability of large video models, such as LTX-Video-13B and Wan2.1-14B.

Training strategy. Table 5 conducts an ablation study to validate our choice of training strategy for improving consistency. The results indicate that supervised finetuning (SFT) demands substantially more data to achieve strong performance, while GRPO performs well even with limited data, making it better suited for the data-scarce shared world modeling task. Furthermore, GRPO alone achieves performance comparable to the combined SFT+GRPO setup, justifying the use of GRPO-only fine-tuning to reduce training cost while maintaining effectiveness.

Reward models. Table 6 and Figure 5 conduct an ablation study on the two proposed reward models. Compared with w/o geometry and motion rewards (first row), introducing the geometry reward (second row) notably improves geometry consistency (e.g., $\text{Geometry}_{0.5}$: 0.7270 vs. 0.7180) while reducing M-FID, showing that geometry consistency benefits from our explicit 3D reconstruction reward. Similarly, adding motion reward (third row) achieves better motion consistency (e.g., Motion_{20} : 0.8445 vs. 0.8364). When both geometry and motion

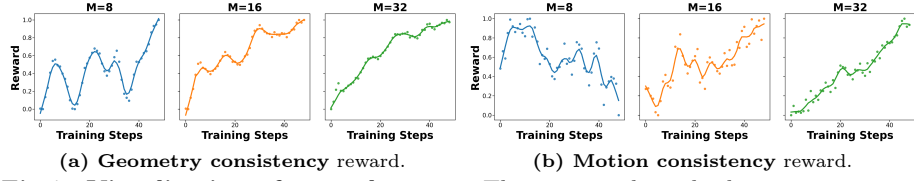


Fig. 7: Visualization of reward curves. The generated results become more consistent with the rewards increase, and the larger group size M leads to a more stable optimization process.

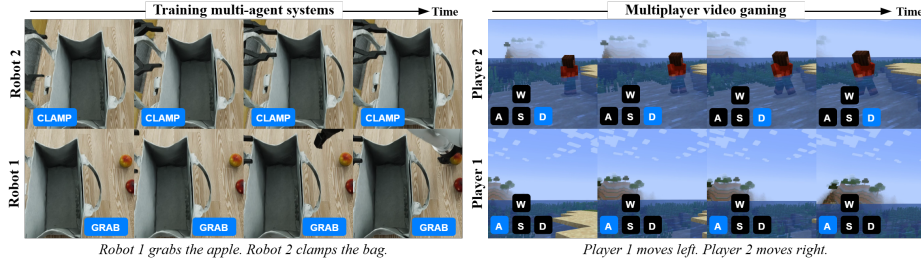


Fig. 8: Applications of IC-World. Our method can facilitate multiple advanced applications. **Left:** Two collaborative robots share the same workspace and coordinate to place an apple into a bag. **Right:** Two players explore the shared world in a first-person video game, where Player 1 leads while Player 2 follows.

rewards are combined (forth row), IC-World achieves consistent improvements across several metrics while maintaining comparable performance on the rest. This demonstrates the general adaptability of the two proposed reward models for enhancing both geometry and motion consistency. Furthermore, in Appendix 6 we analyze the performance fluctuation caused by the external models, and the results show that training becomes less sensitive to reward fluctuation with increasing M .

4.4 Further Analysis

Visualizing the reward curves. As shown in Figure 7, with the group size M increasing from 8 to 32, the reward curves become smoother and more stable, indicating reduced variance and improved convergence. Larger groups provide more reliable relative advantage estimation, leading to steadier optimization during GRPO fine-tuning.

Applications. Beyond conventional applications like multi-view video generation, our framework enables high-level semantic control through instructions, facilitating synchronized generation for advanced applications such as training multi-agent systems and synchronized generation in multiplayer gaming, as shown in Figure 8.

Limitations. The pixel-wise coupling operation for N input images in our in-context generation framework inevitably reduces the effective resolution of each input, resulting in lower-resolution video outputs. A practical workaround is to

apply video super-resolution techniques, such as Upscale-a-Video [51]. Another limitation arises from the inherent capacity of current foundation models: as N increases, maintaining temporal stability and cross-view coherence becomes challenging, often leading to degraded or inconsistent outputs. See Appendix [10] for details.

5 Conclusion

In this paper, we introduced IC-World, a novel framework for shared world modeling built upon the inherent in-context generation capability of large video models. Unlike conventional video-based world modeling pipelines that generate views independently, IC-World enables parallel and coherent generation through pixel-wise coupling and instruction-based in-context prompting. To further enhance geometry and motion consistency, we proposed two complementary reward models built upon 3D reconstruction point clouds and 3D point trackings to optimize the framework via GRPO. Our evaluations on newly designed benchmarks empirically demonstrate that our method substantially improves both geometry and motion consistency across the output set of videos while achieving higher efficiency than existing baselines. Beyond surpassing prior approaches, IC-World opens new possibilities for advanced applications in various areas. We hope this work could lay a solid foundation and inspire future research into shared world modeling.

Acknowledgements

This research is supported by the MoE AcRF Tier 2 grant (MOE-T2EP20223-0001) and the MoE AcRF Tier 1 grant (RG14/22). This research is also supported by a Tencent research grant (04IDS001541N022).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bai, S., Li, M., Liu, Y., Tang, J., Zhang, H., Sun, L., Chu, X., Tang, Y.: Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. arXiv preprint arXiv:2505.14231 (2025)

5. Chai, Q., Zheng, Z., Ren, J., Ye, D., Lin, Z., Wang, H.: Causalmace: Causality empowered multi-agents in minecraft cooperative tasks. arXiv preprint arXiv:2508.18797 (2025)
6. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
7. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
8. Chu, Z., Zhang, L., Sun, Y., Xue, S., Wang, Z., Qin, Z., Ren, K.: Sora detector: A unified hallucination detection for large text-to-video models. arXiv preprint arXiv:2405.04180 (2024)
9. Contributors, L.: Lightx2v: Light video generation inference framework. <https://github.com/ModelTC/lightx2v> (2025)
10. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
11. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639 (2025)
12. Fei, Z., Qiu, D., Li, D., Yu, C., Fan, M.: Video diffusion transformers are in-context learners. arXiv preprint arXiv:2412.10783 (2024)
13. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
14. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
15. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
16. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
17. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
19. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *ICLR* **1**(2), 3 (2022)
21. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)

22. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
23. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
24. Kuang, Z., Cai, S., He, H., Xu, Y., Li, H., Guibas, L., Wetzstein, G.: Collaborative video diffusion: Consistent multi-video generation with camera control. In: arXiv (2024)
25. Li, J., Feng, W., Fu, T.J., Wang, X., Basu, S., Chen, W., Wang, W.Y.: T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *Advances in neural information processing systems* **37**, 75692–75726 (2024)
26. Li, Y., Harada, T.: Leopard: Learning partial point cloud matching in rigid and deformable scenes. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022)
27. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
28. Liu, M., Fang, Z., Li, X., Buhmann, J.M., Ye, D., Li, X., Loy, C.C.: Point-in-context: Understanding point cloud via in-context learning: M. liu et al. *International Journal of Computer Vision* **134**(5), 236 (2026)
29. Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8009–8019 (2025)
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
31. Qwen-Image, T.: Qwen-image technical report (2025), <https://arxiv.org/abs/2508.02324>
32. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
33. Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., Chadha, A.: A comprehensive survey of hallucination in large language, image, video and audio foundation models. arXiv preprint arXiv:2405.09589 (2024)
34. Wan, T.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, C., Zhuang, P., Ngo, T.D., Menapace, W., Siarohin, A., Vasilkovsky, M., Skorokhodov, I., Tulyakov, S., Wonka, P., Lee, H.Y.: 4real-video: Learning generalizable photo-realistic 4d video diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17723–17732 (2025)
36. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)
37. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3 : Scalable permutation-equivariant visual geometry learning (2025), <https://arxiv.org/abs/2507.13347>

38. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26057–26068 (2025)
39. Wu, W., Guo, X., Tang, W., Huang, T., Wang, C., Chen, D., Ding, C.: Drivescape: Towards high-resolution controllable multi-view driving video generation. arXiv preprint arXiv:2409.05463 (2024)
40. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3d point tracking made easy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025), <https://arxiv.org/abs/2507.12462>
41. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. arXiv preprint arXiv:2407.17470 (2024)
42. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Wang, X., Wong, T.T., Shan, Y.: Dynamicrafter: Animating open-domain images with video diffusion priors. arXiv preprint arXiv:2310.12190 (2023)
43. Xu, J., Zou, X., Huang, K., Chen, Y., Liu, B., Cheng, M., Shi, X., Huang, J.: Easyanimate: A high-performance long video generation method based on transformer architecture. arXiv preprint arXiv:2405.18991 (2024)
44. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrp: Unleashing grp on visual generation. arXiv preprint arXiv:2505.07818 (2025)
45. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
46. Yuan, H., Chen, Z., Ji, K., Gu, Q.: Self-play fine-tuning of diffusion models for text-to-image generation. *Advances in Neural Information Processing Systems* **37**, 73366–73398 (2024)
47. Zhang, J., Wu, J., Chen, W., Ji, Y., Xiao, X., Huang, W., Han, K.: Onlinevpo: Align video diffusion model with online video-centric preference optimization. arXiv preprint arXiv:2412.15159 (2024)
48. Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690 (2025)
49. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 10412–10420 (2025)
50. Zheng, W., Song, R., Guo, X., Zhang, C., Chen, L.: Genad: Generative end-to-end autonomous driving. In: European Conference on Computer Vision. pp. 87–104. Springer (2024)
51. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2535–2545 (2024)