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

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

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

[*] Equal contribution [†] Work done during an internship at MSRA [‡] Corresponding authors

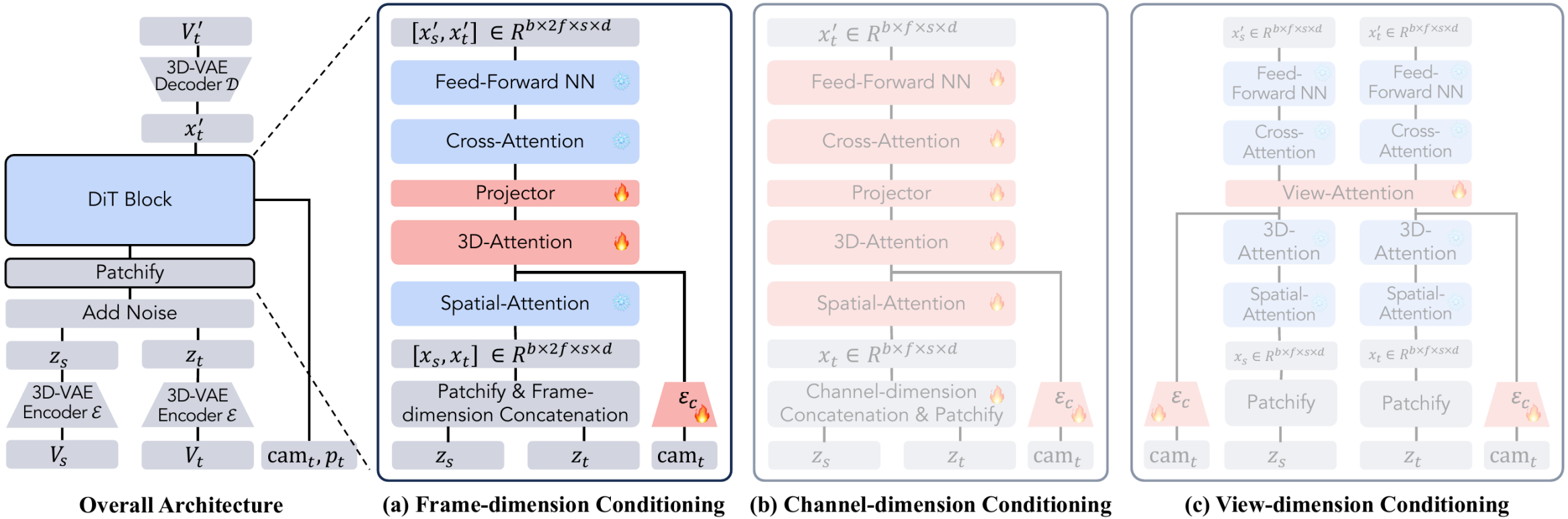
Video Foundation Models

Prompt: Camera move left. Modernist glass skyscrapers reflecting the Shanghai Bund waterfront during golden hour.

Poor controllability

Poor 3D consistency

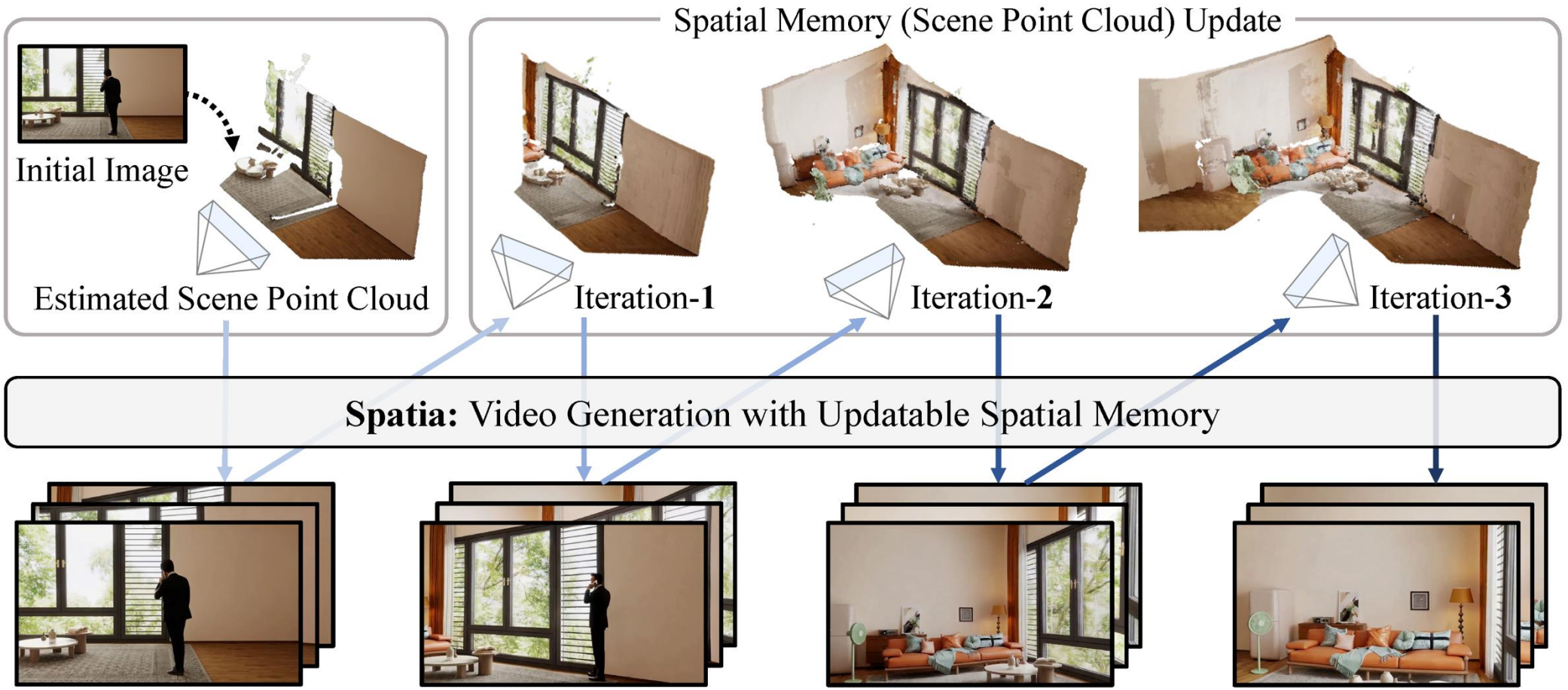
Existing Methods for Camera Control



Overall Architecture **(a) Frame-dimension Conditioning** **(b) Channel-dimension Conditioning** **(c) View-dimension Conditioning**

New Modules for Controller

Existing Methods for 3D Consistency



Explicit 3D Memory Cache

Motivation

No Explicit Memory Module

No Control Module

} No architectural modifications

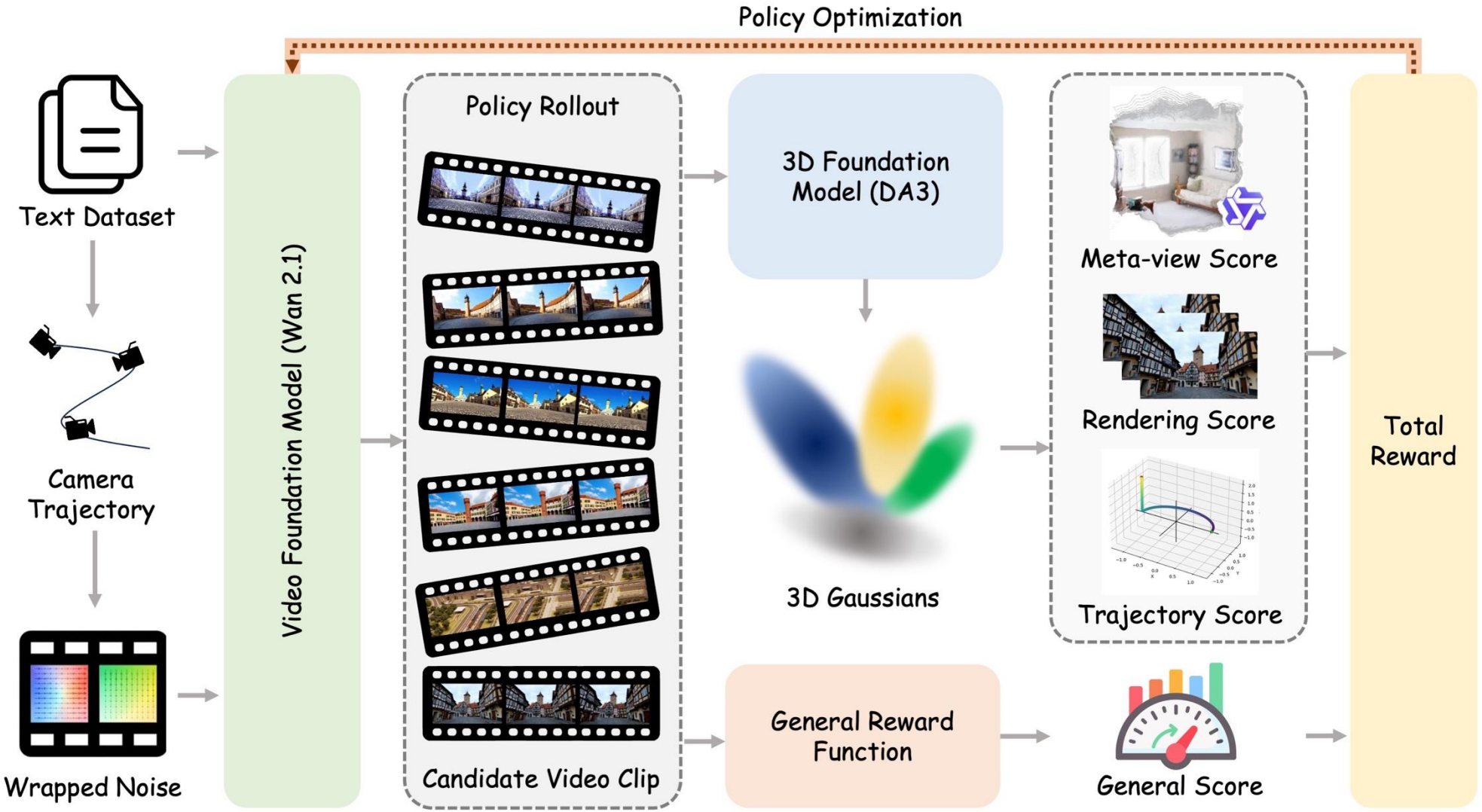
No Specific Video Training Data

No extra inference cost

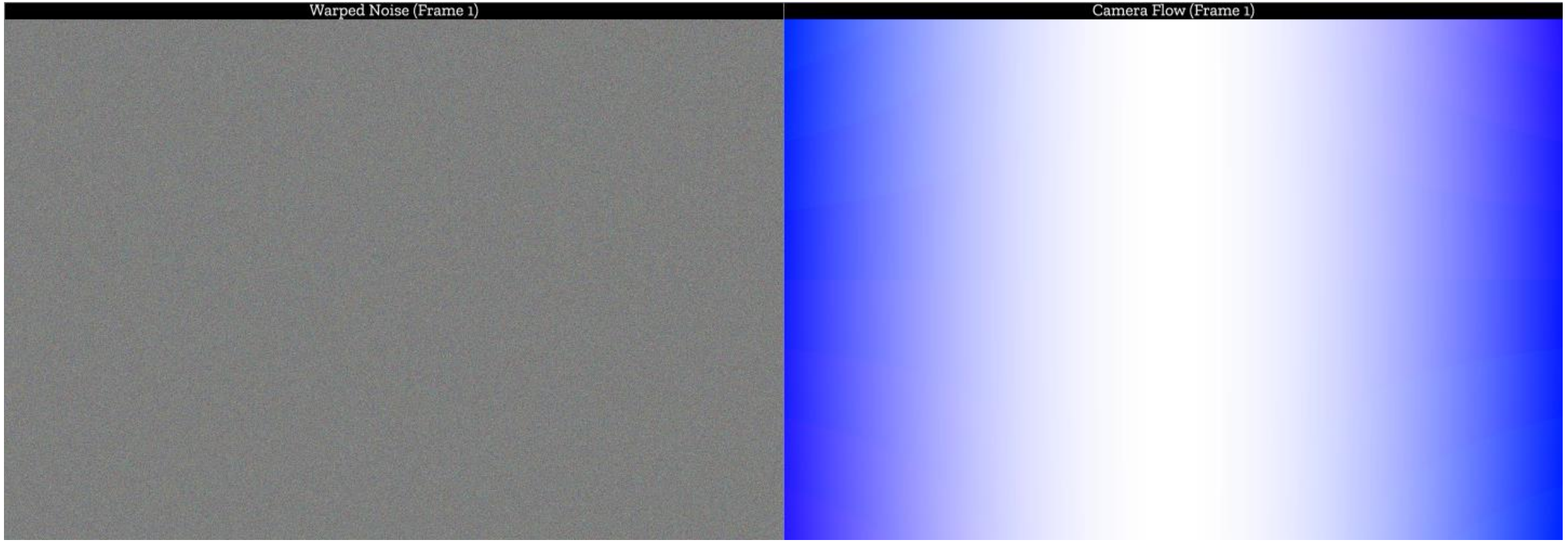
Not only applicable to I2V

ALL-IN-ONE-METHOD

Overview

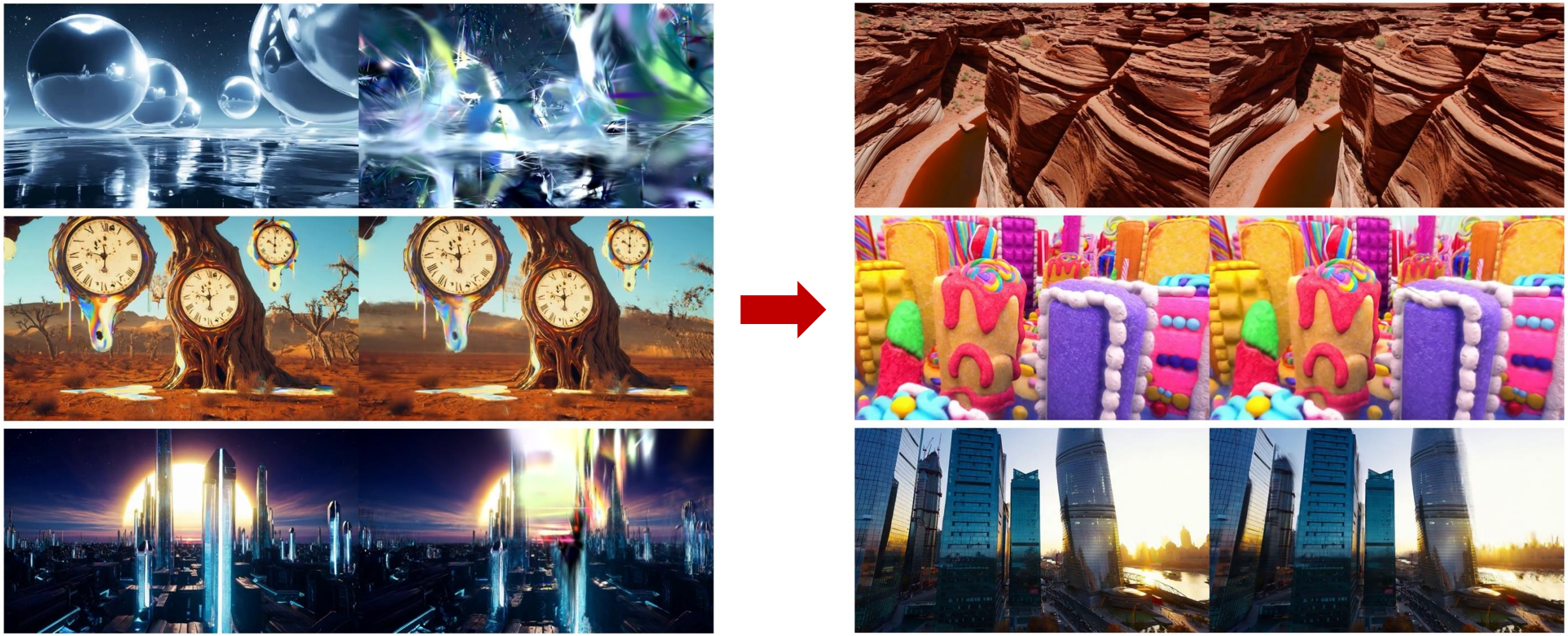

Policy Optimization
Text Dataset
Camera Trajectory
Wrapped Noise
Video Foundation Model (Wan 2.1)
Policy Rollout
Candidate Video Clip
3D Foundation Model (DA3)
3D Gaussians
General Reward Function
Meta-view Score
Rendering Score
Trajectory Score
General Score
Total Reward

Implicit Camera Conditioning



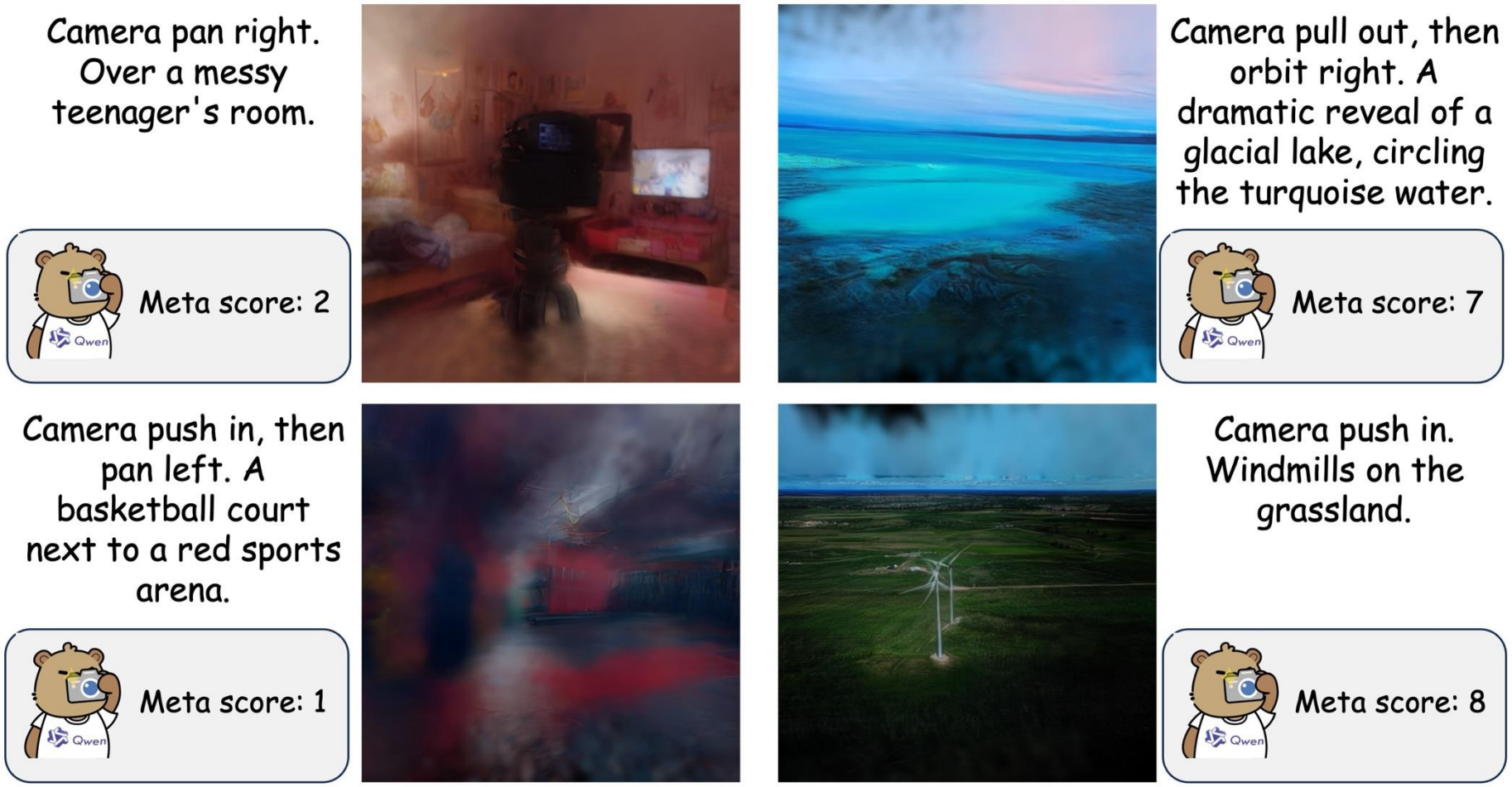
Instead of using Gaussian noise, using warped noise

Reward Design: Rendering Score



Re-rendered visual alignment metrics

Reward Design: Meta-view Score

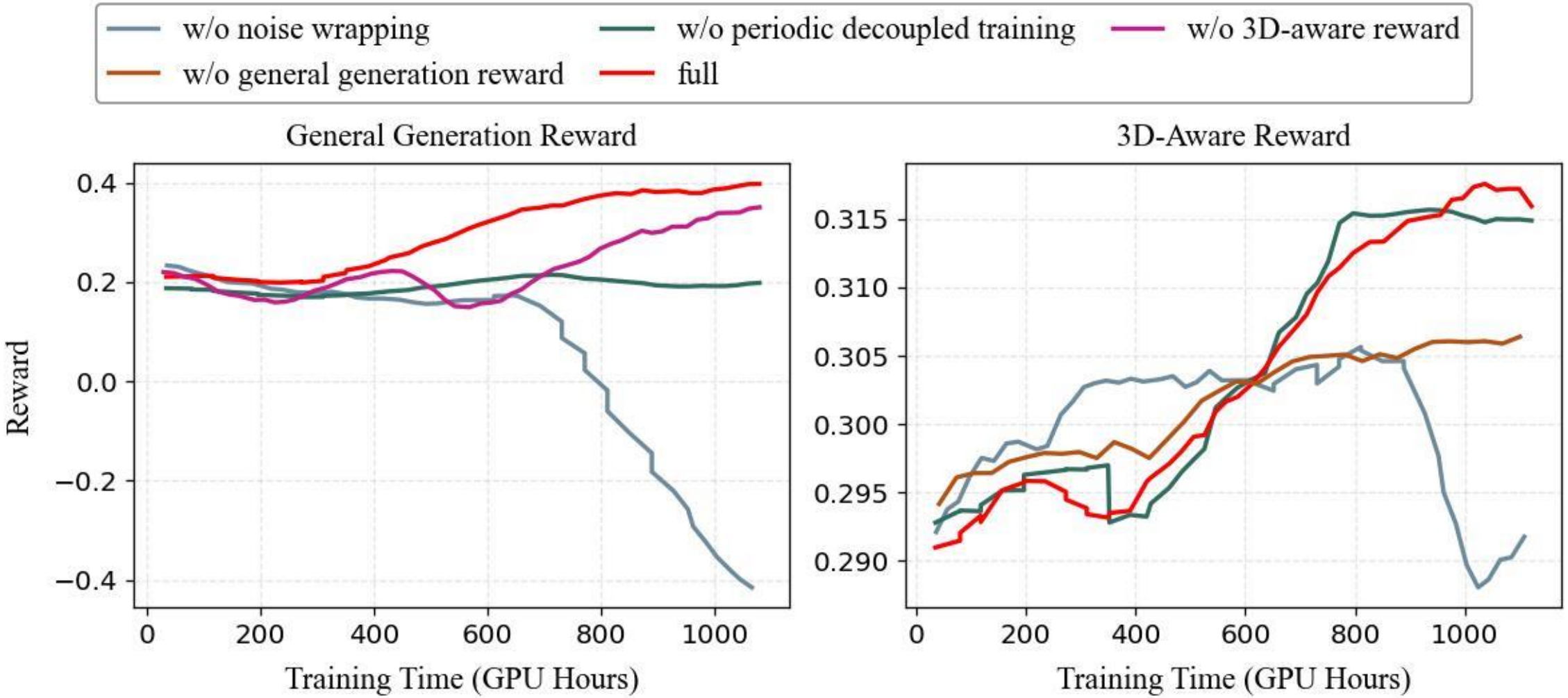


VLM determines the quality of the generated scene

Reward Design: Trajectory Score

Error in calculating the injected camera trajectory

Training Strategy



Periodic Decoupled Training ensures the quality of dynamic scenes.

Quantitative Results

Method	PSNR ↑	SSIM ↑	LPIPS ↓
CogVideoX-1.5-5B (Yang et al., 2025)	24.44	0.783	0.242
Wan2.2-T2V-14B (Wan et al., 2025)	23.47	0.779	0.253
Wan2.2-T2V-5B (Wan et al., 2025)	22.36	0.716	0.303
Wan2.1-T2V-14B (Wan et al., 2025)	19.76	0.629	0.405
Wan2.1-T2V-1.3B (Wan et al., 2025)	17.40	0.550	0.467
World-R1-Small (Ours)	27.63	0.858	0.201
World-R1-Large (Ours)	**27.67**	**0.865**	**0.162**

+10dB PSNR, +0.3 SSIM, -0.3 LPIPS

Quantitative Results

Method	Aesthetic Quality ↑	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Background Consistency ↑
CogVideoX-1.5-5B (Yang et al., 2025)	62.07	65.34	98.15	96.56	96.81
Wan2.1-T2V-1.3B (Wan et al., 2025)	62.43	66.51	97.44	96.34	97.29
GCD (Van Hoorick et al., 2024)	38.21	41.56	98.37	88.94	92.00
Trajectory-Attention (Xiao et al., 2024)	38.50	51.00	98.21	90.60	92.83
DAS (Gu et al., 2025)	39.86	51.55	99.14	90.34	92.03
ReCamMaster (Bai et al., 2025a)	42.70	53.97	99.28	92.05	93.83
World-R1-Small (Ours)	65.74	67.53	98.55	97.58	96.67

Even higher general video generation metrics

Comparation with Baselines

Camera push in. Deep canyon walls made of layered red rock, with a winding river at the bottom.

Wan2.2-T2V-14B

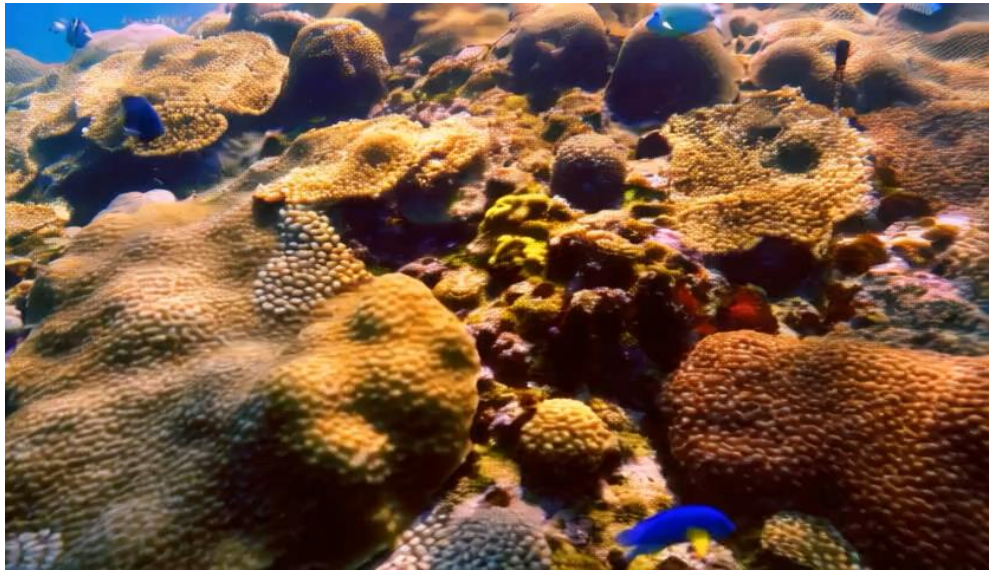
Wan2.1-T2V-14B

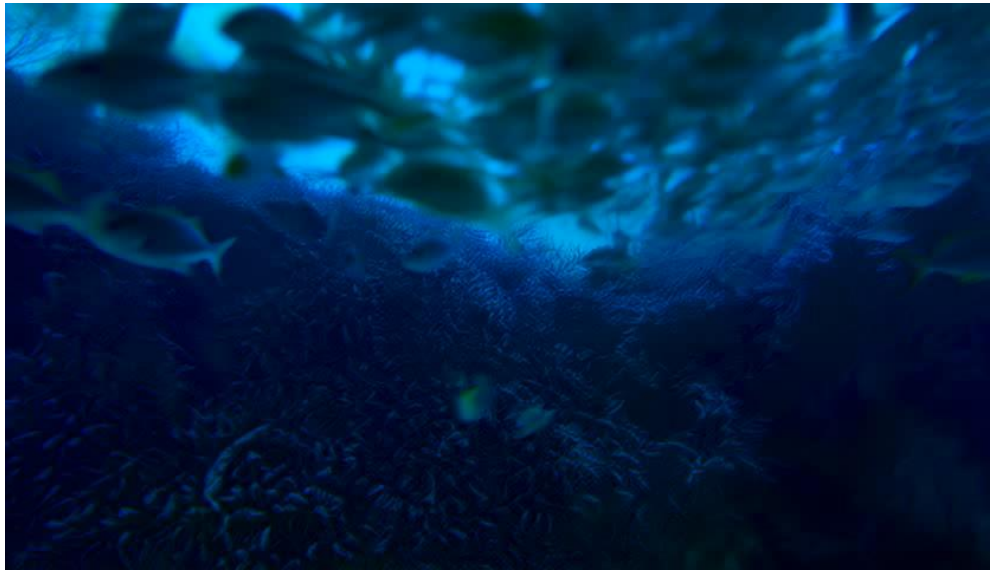
World-R1-Large (Ours)

Comparation with Baselines

Camera orbit left, then push in. Deep sea coral reefs teeming with colorful fish and bioluminescent plant life.

Wan2.2-T2V-14B





Wan2.1-T2V-14B

World-R1-Large (Ours)

Comparation with Baselines

Camera move left. Modernist glass skyscrapers reflecting the Shanghai Bund waterfront during golden hour.

CogVideoX-1.5-5B

Wan2.1-T2V-1.3B

World-R1-Small (Ours)

Results of Dynamic Scene

World-R1-Large (Ours)

A lion roaring with its mane shaking in the wind.

Camera pan right. A drone flying through a complex obstacle course.

Results of Dynamic Scene

World-R1-Large (Ours)

Camera pan left. Soldiers marching in synchronization across a dusty field.

Camera move left. A fighter jet performing an aileron roll.

More Information

Paper, code and more visualizations are available on our project page.

Weijie Wang's homepage. Actively seeking cooperation opportunities.

Conclusion:

- **World-R1** reinforces 3D constraints for text-to-video generation without extra inference-time modules.
- **Improving** camera control and 3D coherence across static and dynamic scenes.
- **Can** be scaled up to the video foundation model.