Towards Intelligent Interactive 3D-aware Video World Model

Weijie Wang

http://lhmd.top

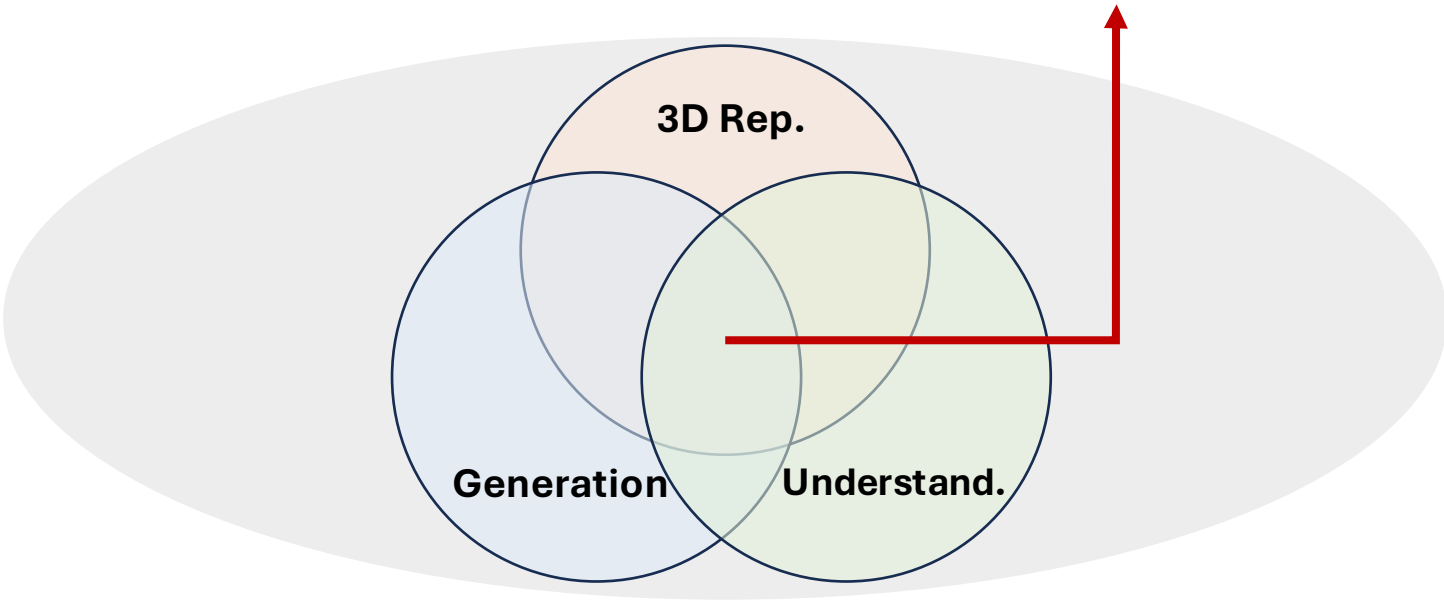
College of Computer Science and Technology, Zhejiang University

Ph.D. student @ ZIP Lab / Intern @ ByteDance Seed

2026/06/17

One Story

Build an intelligent, interactive, 3D-aware world model



Observe
video / images / point cloud

Reconstruct
feed-forward explicit 3D?

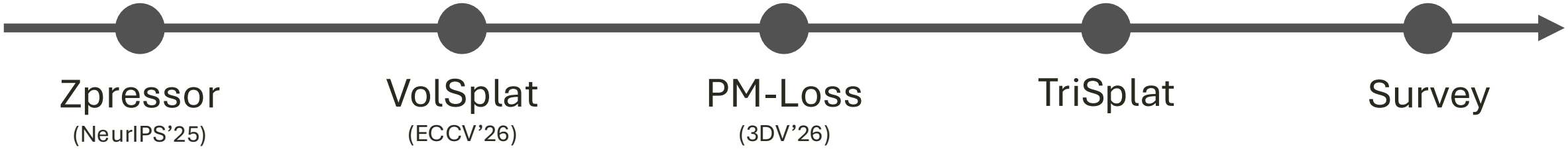
Understand & Generate
language / video / memory

Interact
mesh / physics / simulation

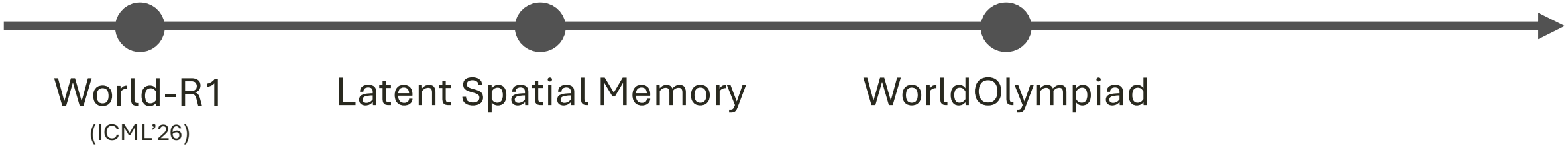
3D is the interface that can connect reconstruction, generation, understanding, and physical interaction.

Two Key Directions

Key Direction 1: Breaking the limits of Feed-Forward 3D Reconstruction



Key Direction 2: How to construct a 3D-aware video world model



Enhance 3D Consistency for Video World Models via Post-Training

Video Foundation Models

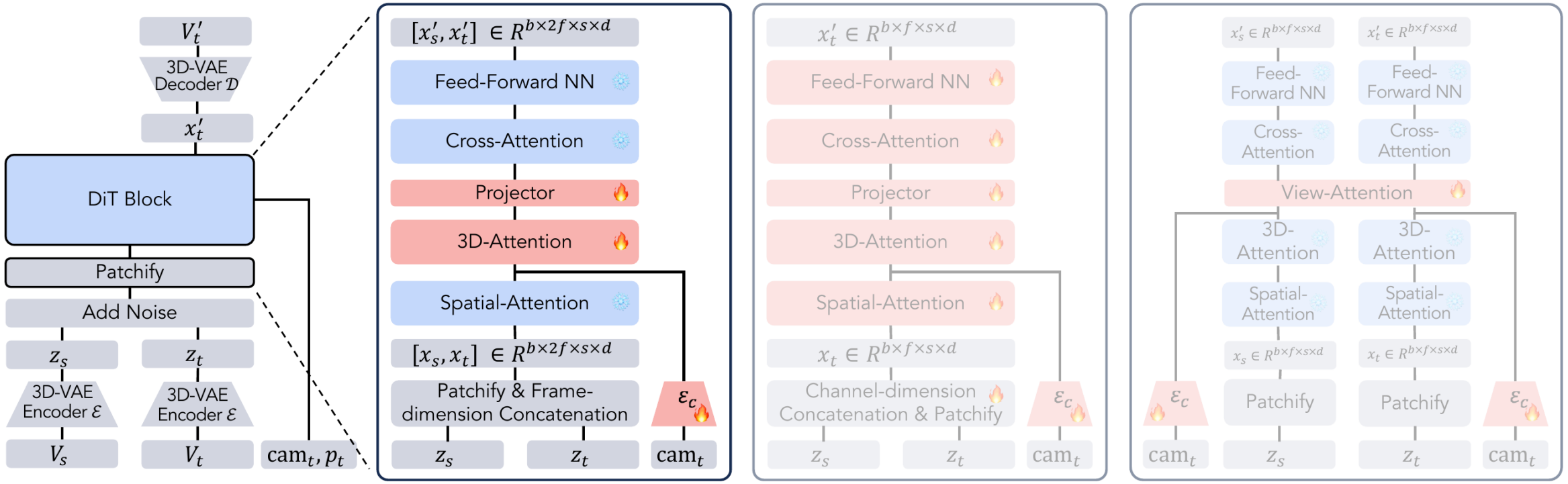
Prompt: Camera move left. Modernist glass skyscrapers reflecting the Shanghai Bund waterfront during golden hour.

Poor controllability

Poor 3D consistency

Existing Methods for Camera Control

Weijie Wang

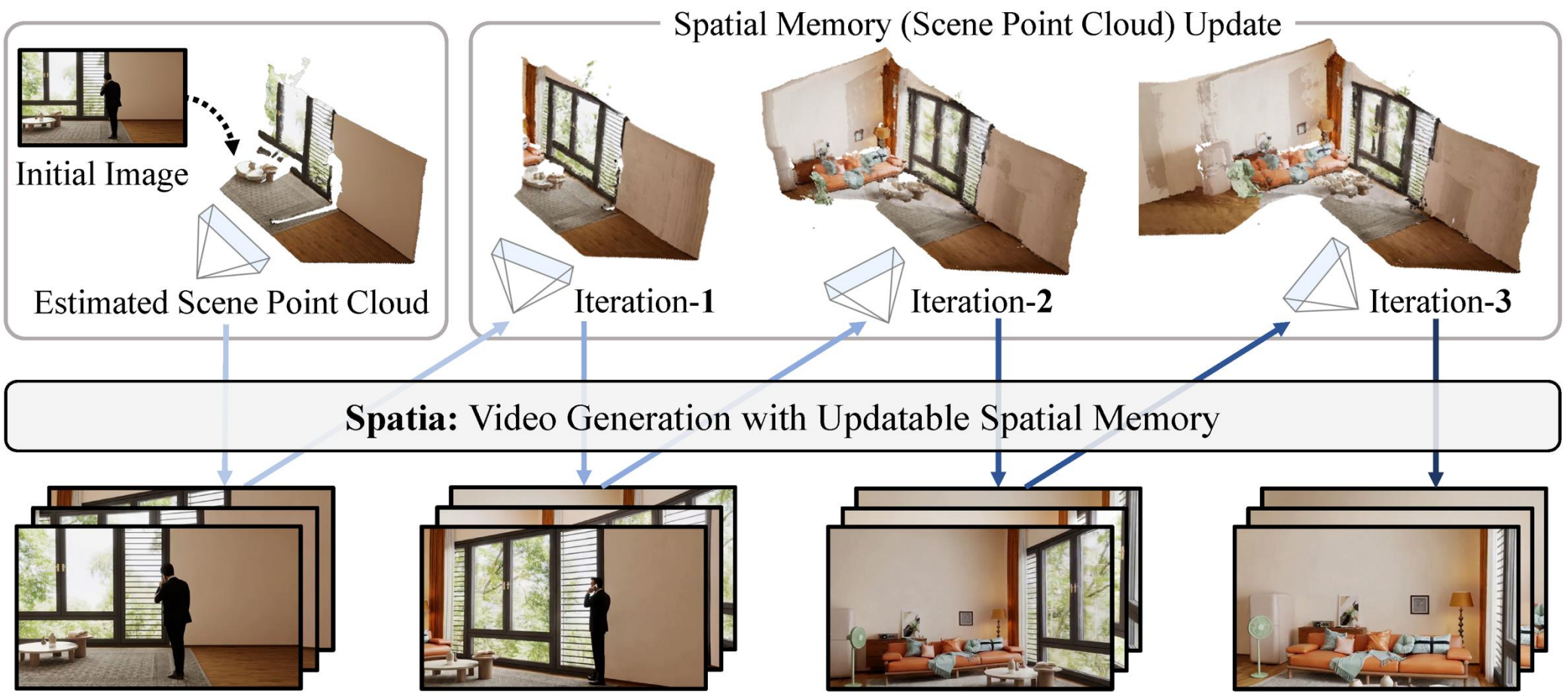


Overall Architecture **(a) Frame-dimension Conditioning** **(b) Channel-dimension Conditioning** **(c) View-dimension Conditioning**

New Modules for Controller

Existing Methods for 3D Consistency

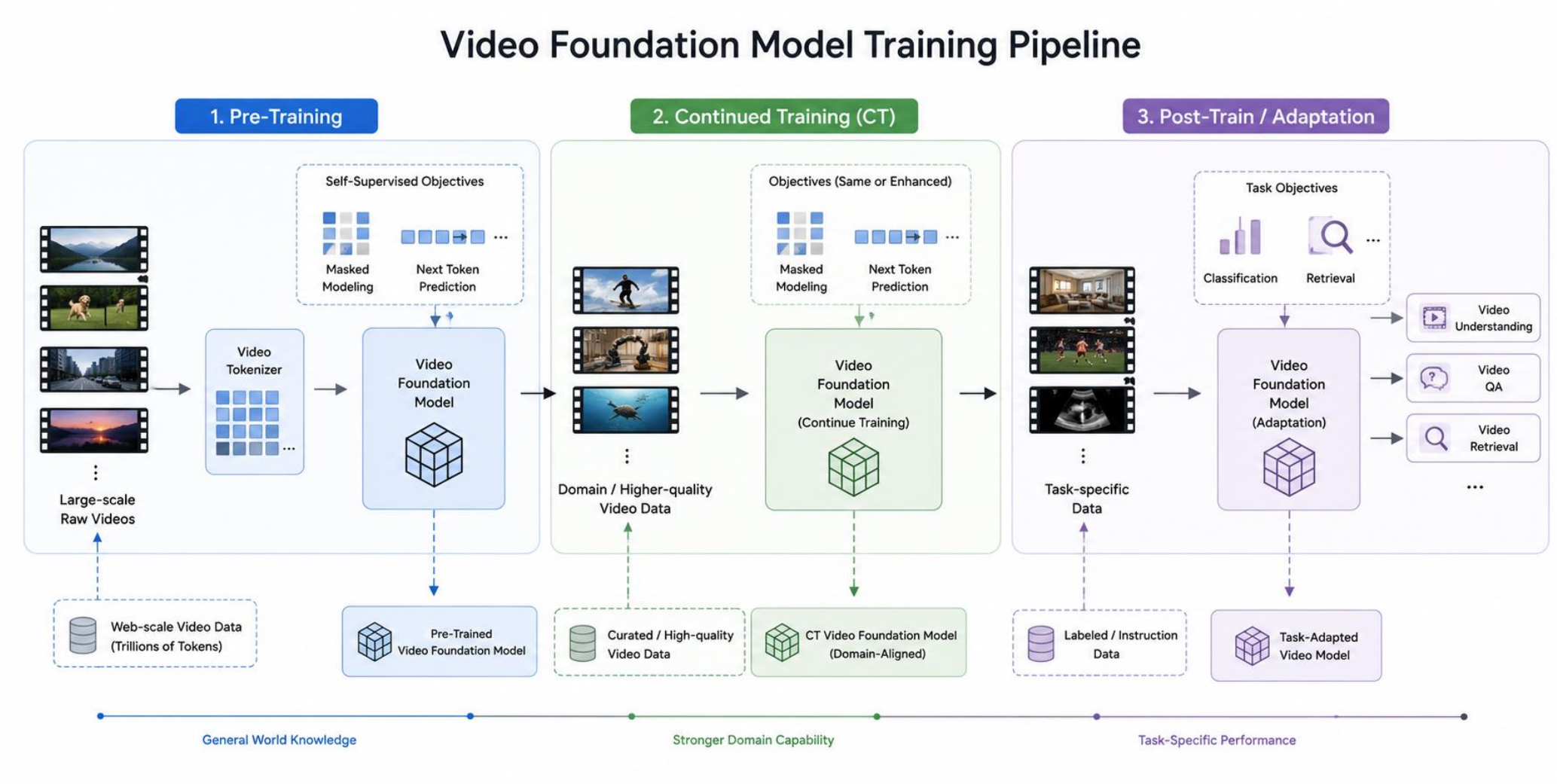
Weijie Wang



Explicit 3D Memory Cache

Video Foundation Models

Weijie Wang



Where to inject 3D information? Pretrain (Data)? CT? or PT?

Motivation

No Explicit Memory Module

No Control Module

} No architectural modifications

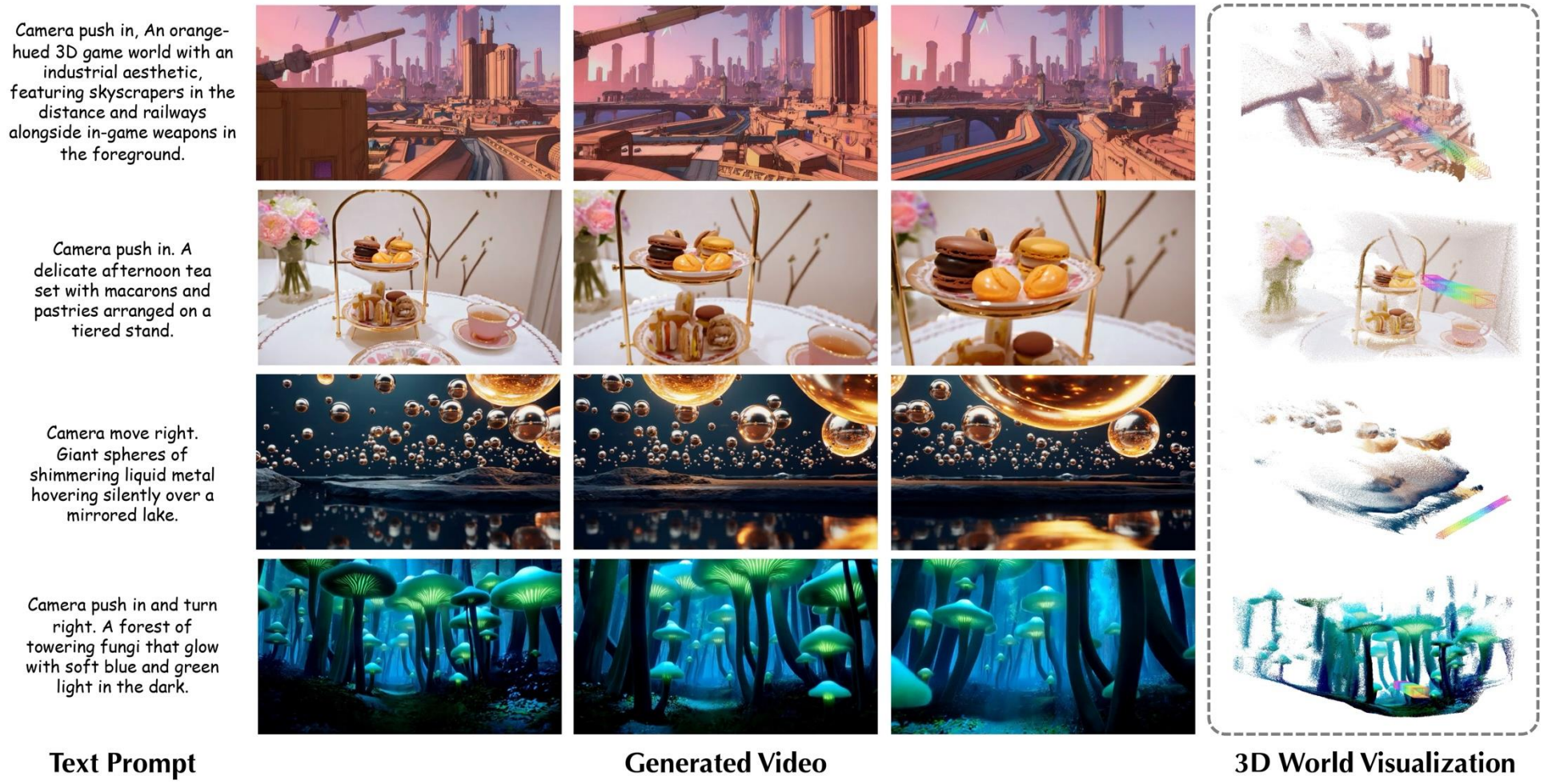
No Specific Video Training Data

No extra inference cost

Not only applicable to I2V

ALL-IN-ONE-METHOD

Our Answer: World-R1 (Post-Train)

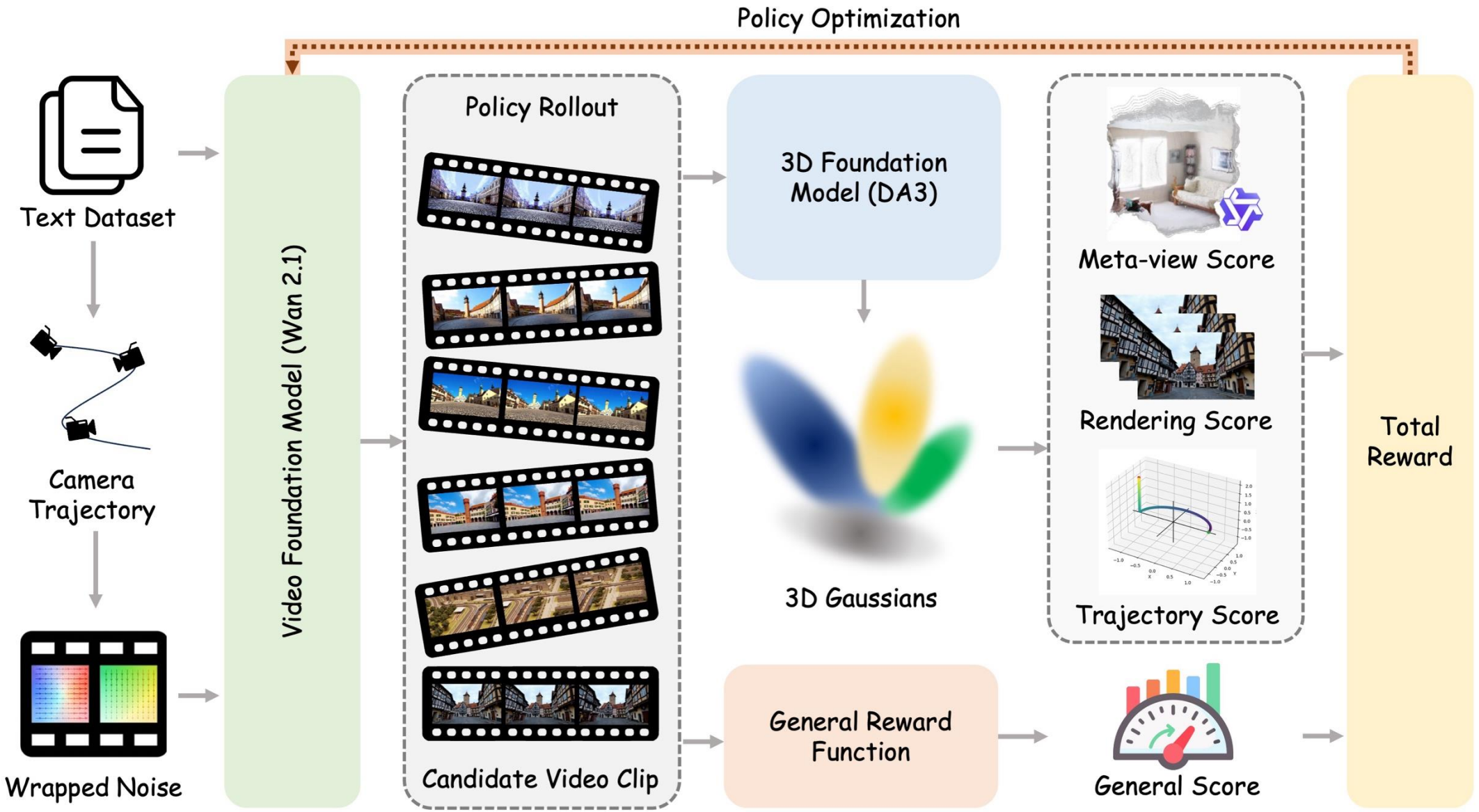


3D consistency can be converted into RL rewards for text-to-video generation.

 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." ICML 2026.

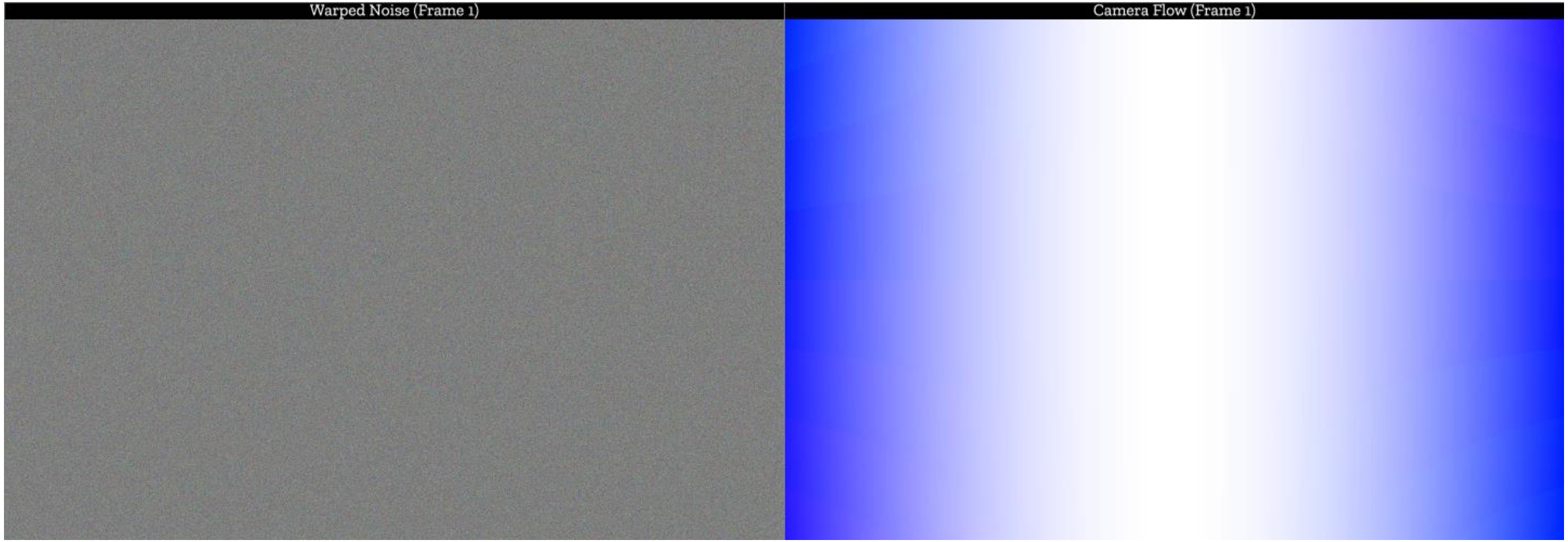
World-R1

Weijie Wang



Implicit Camera Conditioning

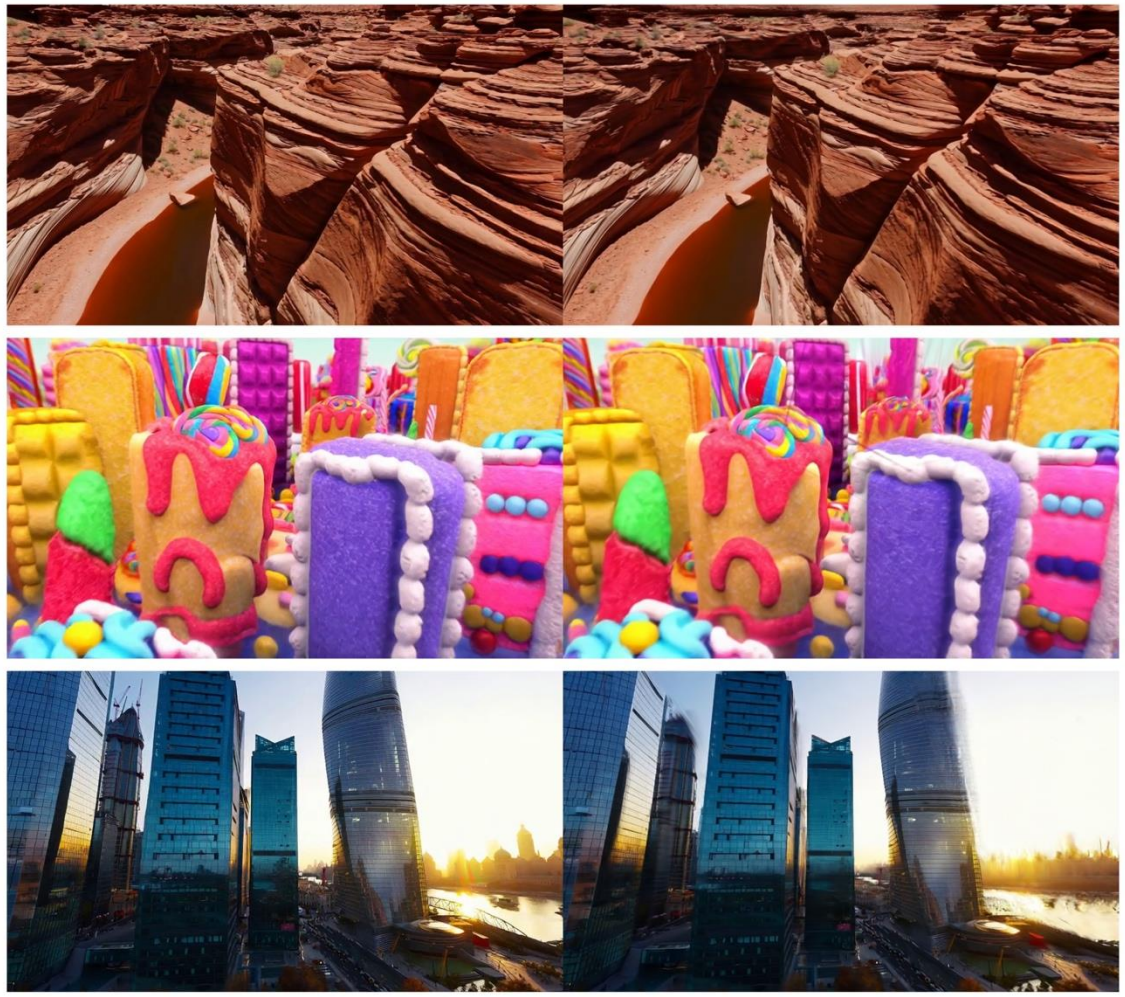
Weijie Wang



Instead of using Gaussian noise, using warped noise

Reward Design: Rendering Score

Generated Video Frame | Reconstructed Video Frame

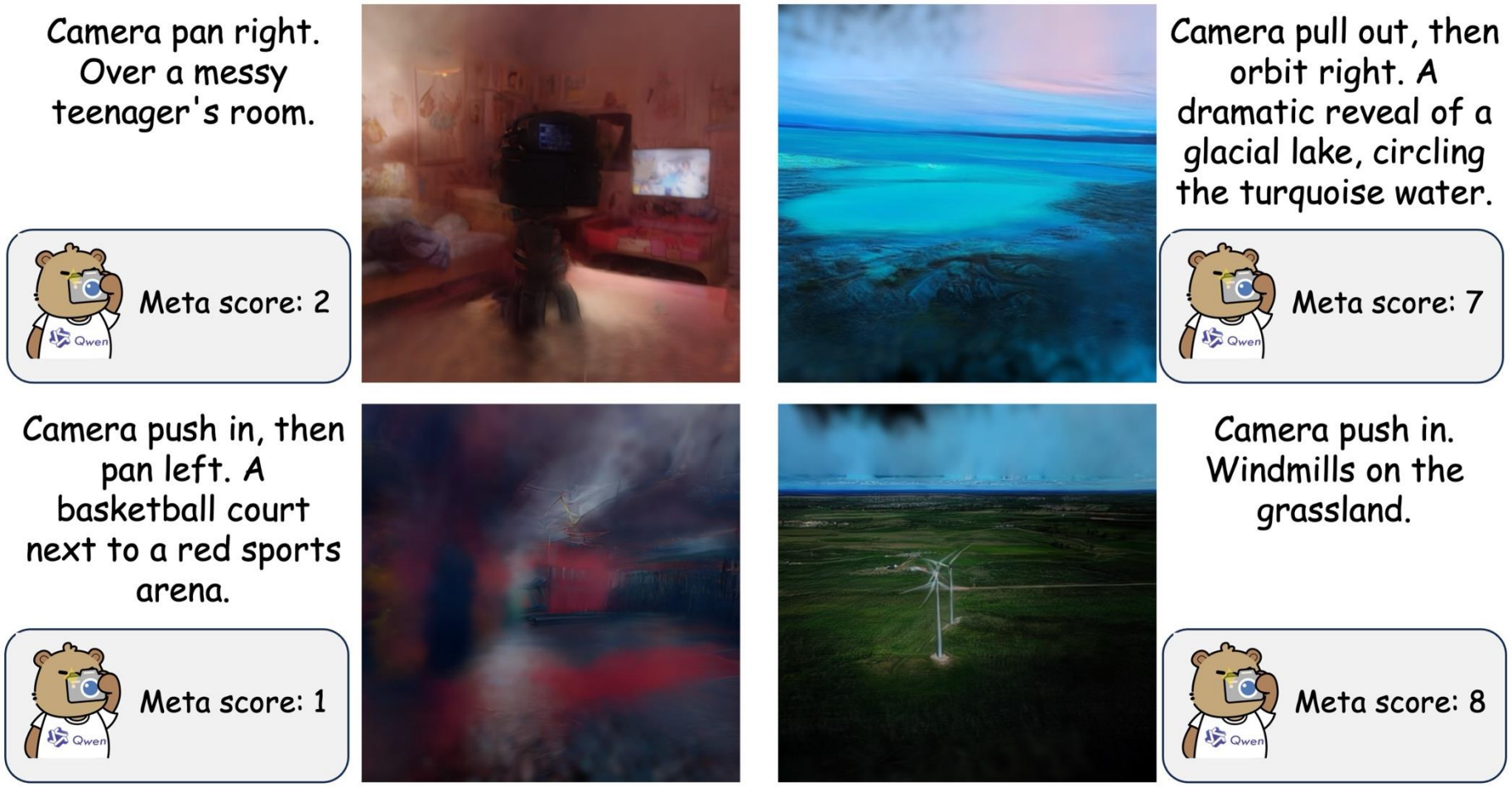


Generated Video Frame | Reconstructed Video Frame

Re-rendered visual alignment metrics

Reward Design: Meta-view Score

Weijie Wang

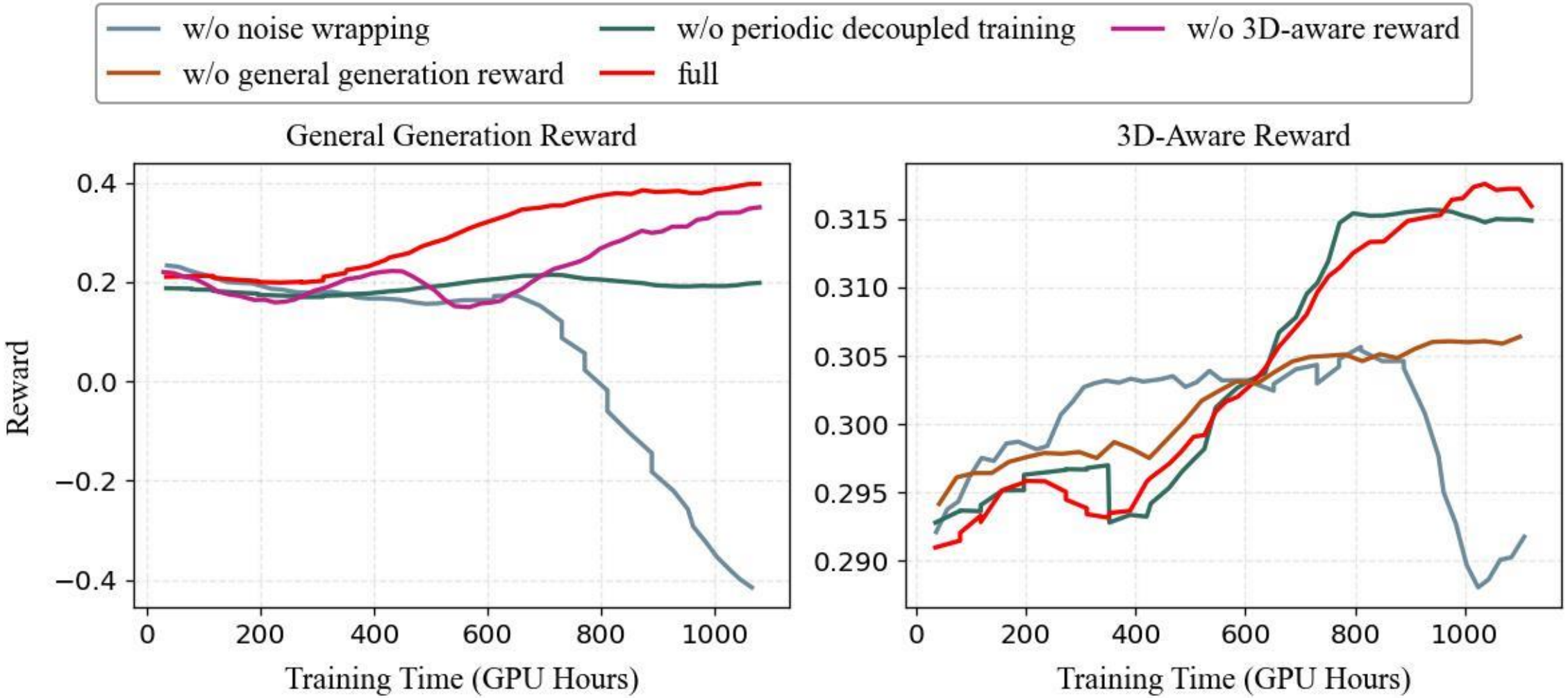


VLM determines the quality of the generated scene

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

Weijie Wang

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 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

Weijie Wang

World-R1-Large (Ours)

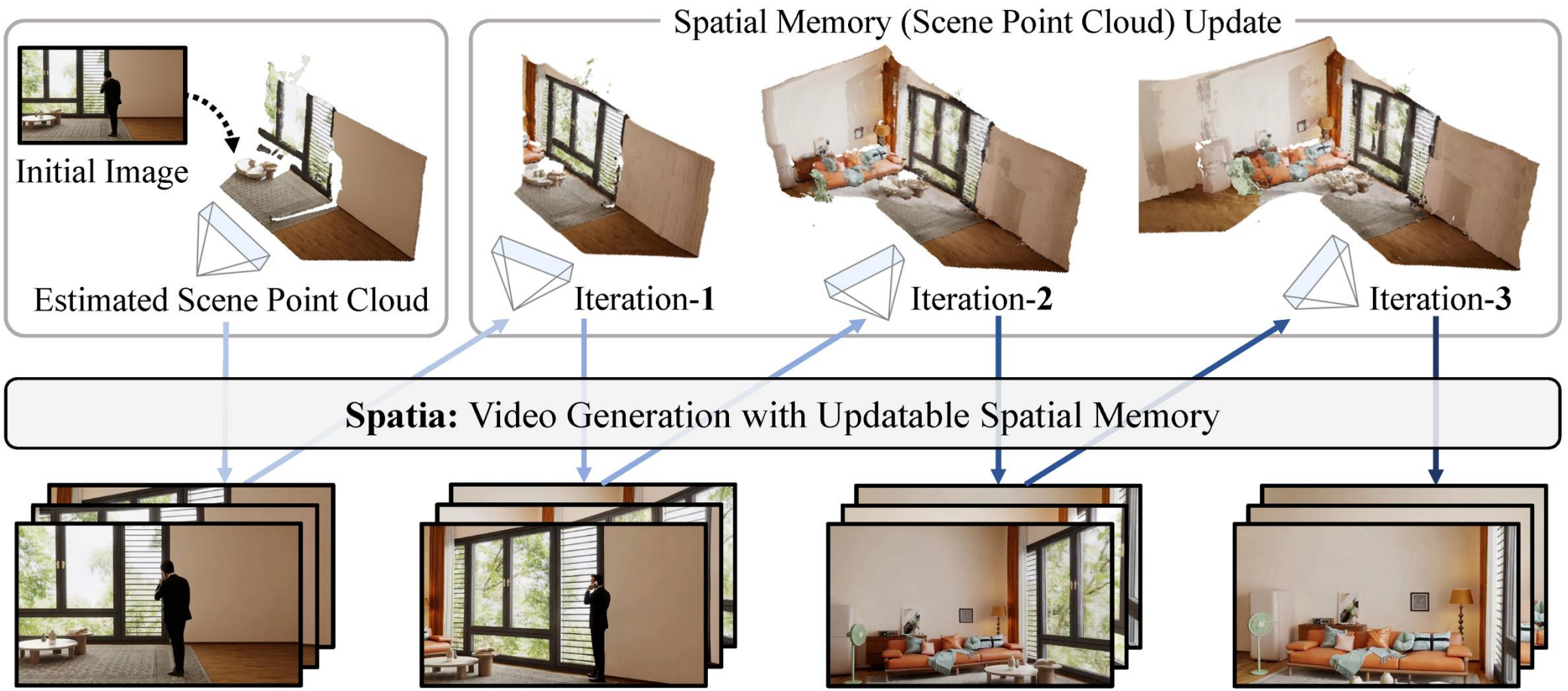
Camera pan left. Soldiers marching in synchronization across a dusty field.

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

But it’s so expensive! We can do much more with 3D memory!

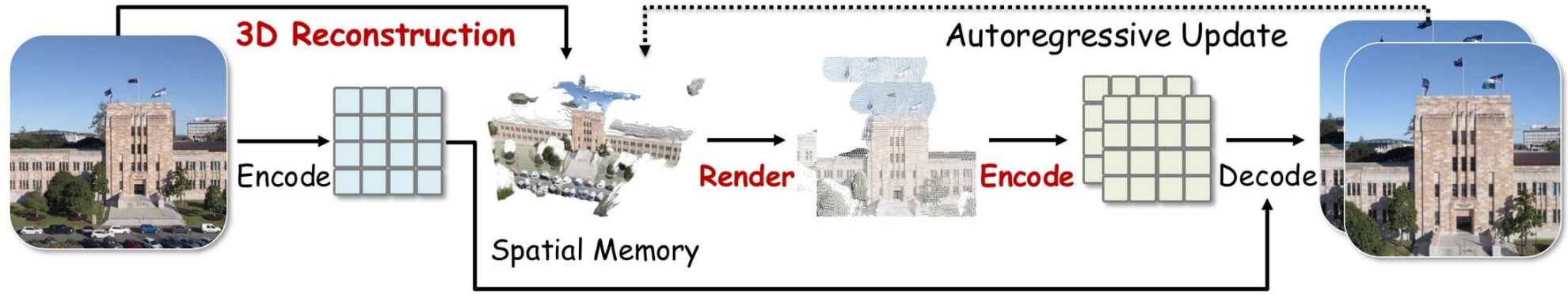
Review: Existing Methods for 3D Consistency

Weijie Wang

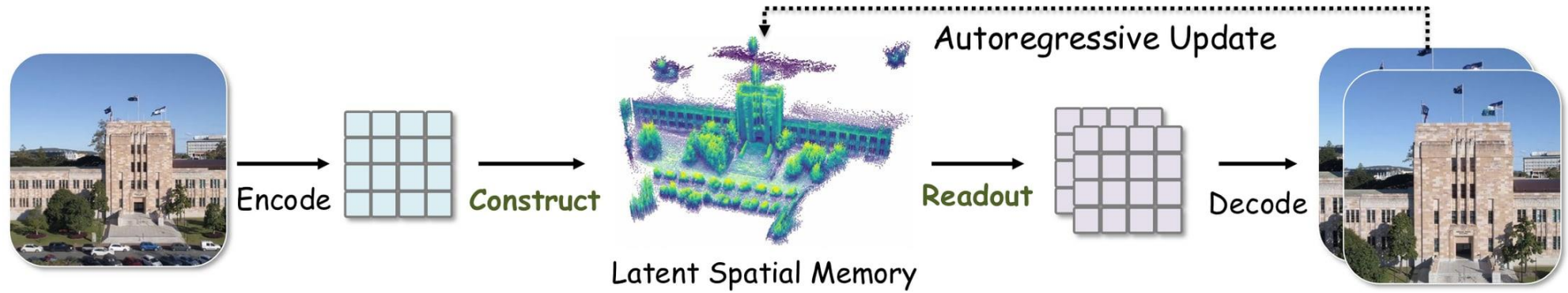


Explicit 3D Memory Cache

Latent Spatial Memory



Spatial Memory for Video World Models



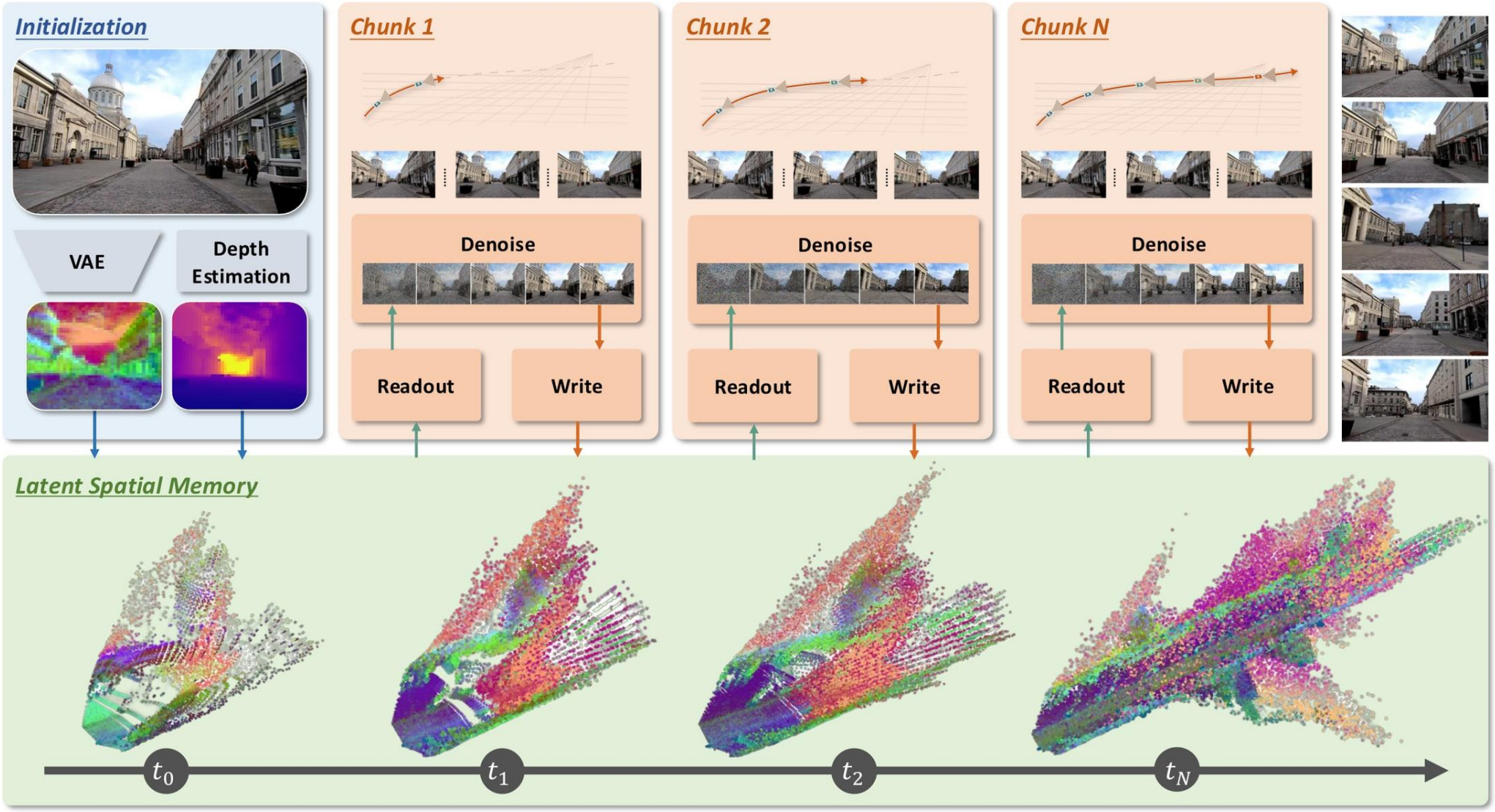
Latent Spatial Memory for Video World Models (Ours)

Video World Models can use Latent Spatial Memory to maintain 3D Consistency

 Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. "Latent Spatial Memory for Video World Models." arXiv 2026.

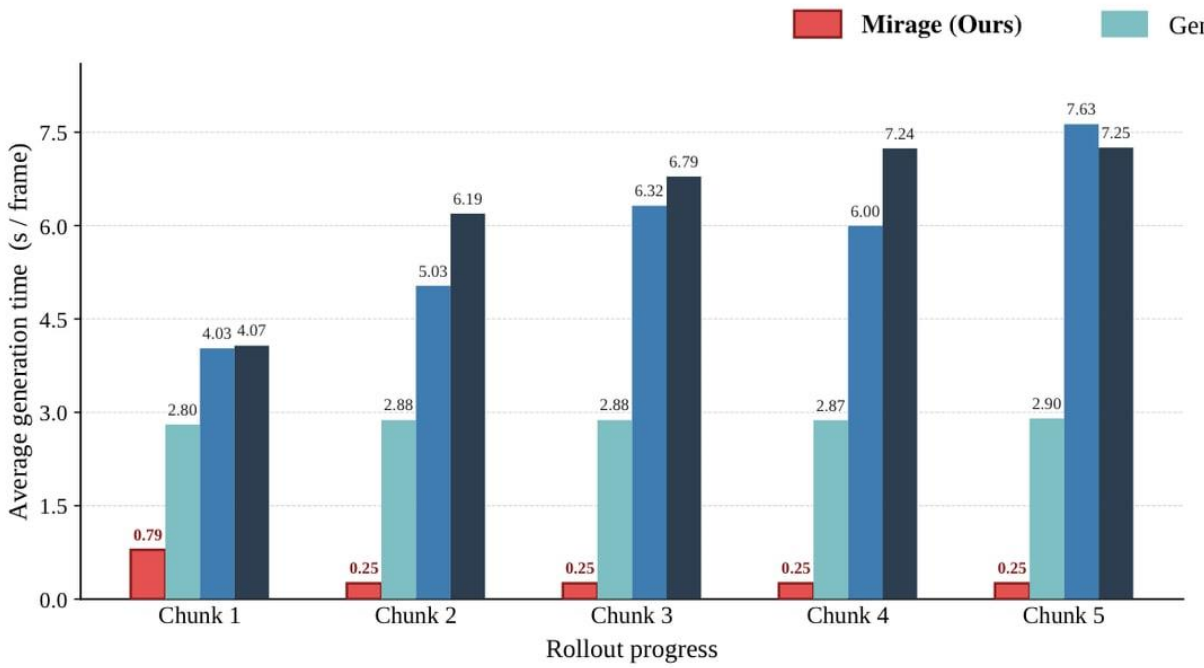
Mirage: Pipeline with Latent Spatial Memory

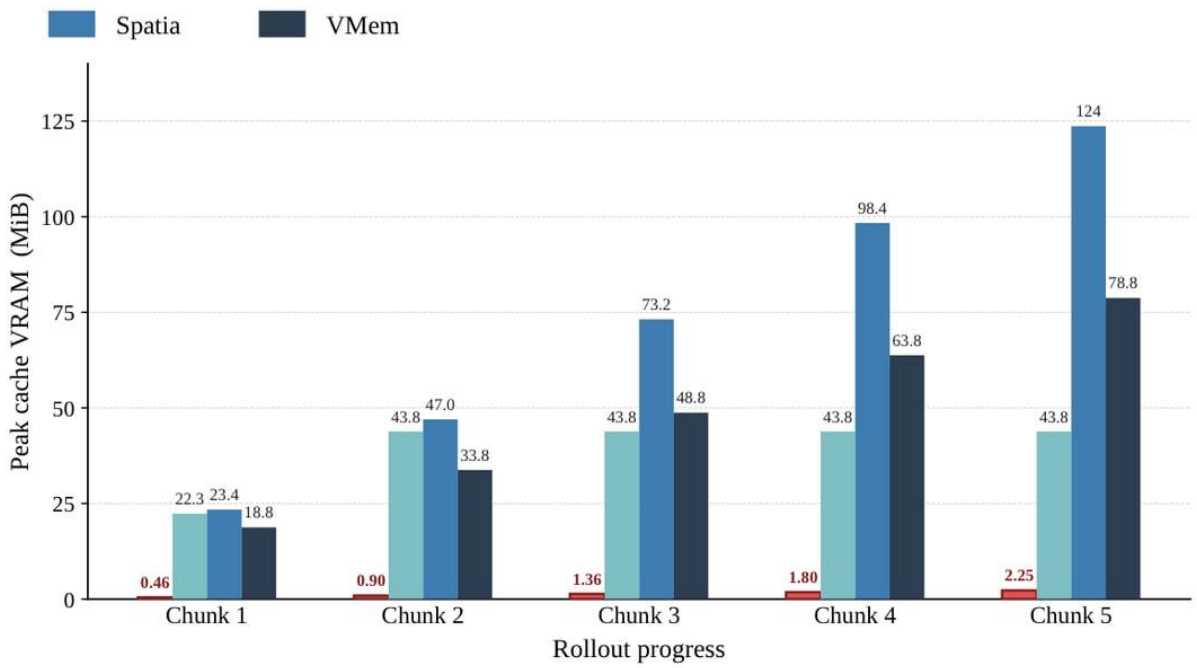
Weijie Wang



Latent Spatial Memory

Weijie Wang





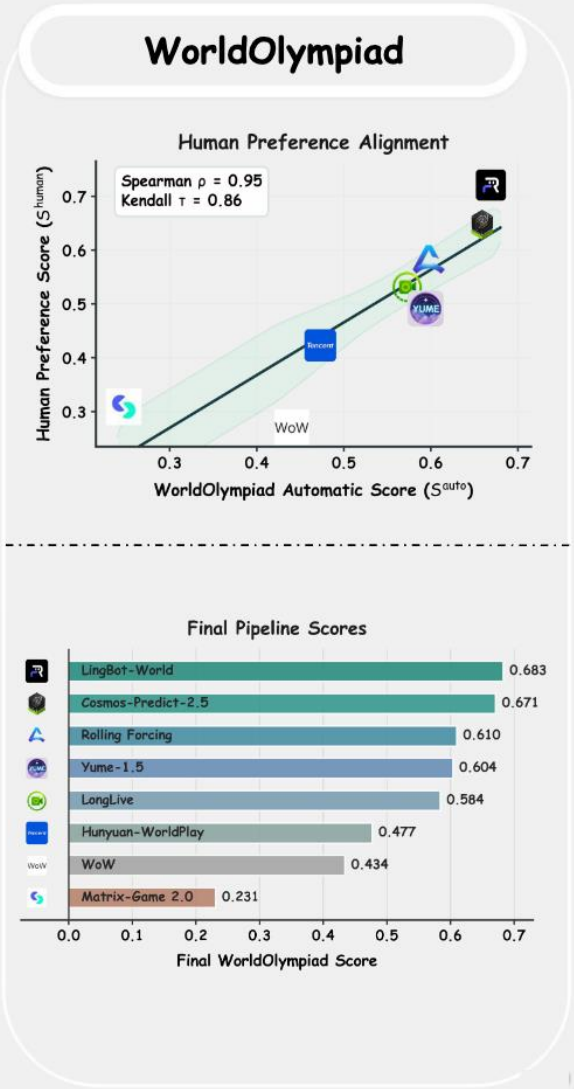
Latent Spatial Memory

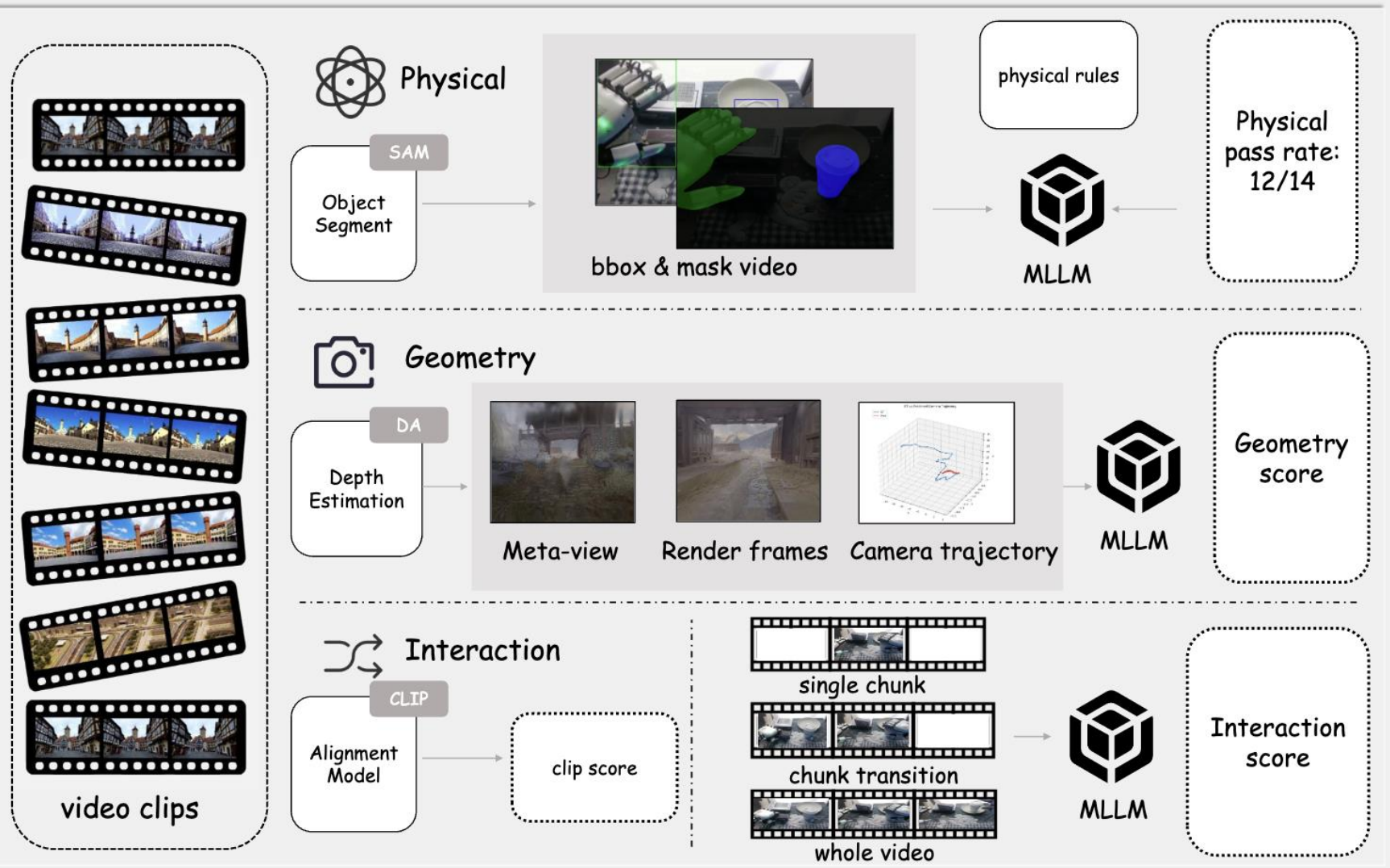
for Video World Models

How to Benchmark these works?

WorldOlympiad

Weijie Wang





WorldOlympiad

Benchmark	Eval Metrics				Video Tasks		
	Long Video	Physical	Geometry	Interaction	Gaming	Robotics	Real-world
VBench [15]	✗	✗	✗	✗	✗	✗	✓
VBench++ [16]	✓	✗	✗	✗	✗	✗	✓
VBench 2.0 [53]	✗	✓	✗	✗	✗	✗	✓
MIND [47]	✓	✗	✗	✓	✓	✗	✗
EWMBench [12]	✗	✗	✗	✓	✗	✓	✗
WorldEval [20]	✗	✗	✗	✓	✗	✓	✗
WorldArena [29]	✗	✓	✓	✓	✗	✓	✗
WorldOlympiad	✓	✓	✓	✓	✓	✓	✓

Summary

- Enhance the 3D capabilities of video foundation models – ***World-R1***
- Use 3D latent to represent the entire world – ***Latent Spatial Memory***
- Benchmarking the video world model – ***WorldOlympiad***

Thank You

Towards Intelligent Interactive 3D-aware Video World Model