Physics-Grounded World Models: Generation, Interaction, and Evaluation

Hong-Xing “Koven” Yu

Stanford University

Video Generation Today: Stunning Visuals

Generated by Veo 3

Pixel Generation as World Models: Challenges

World models must understand how worlds evolve under actions, enabling agents to interact meaningfully.

Pixel generation models (e.g., VideoGen) are bottlenecked by:

❌ **Precise Action Control**

“Push cup to right by 10cm”

❌ **Physical Consistency**

❌ **Efficiency**

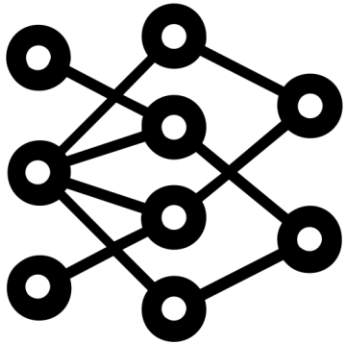
Beyond Scaling: Physics Grounding

Simply scaling up?

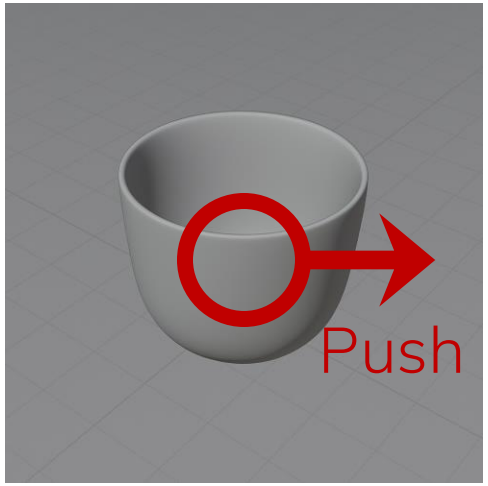
? Precise Action-Video Data

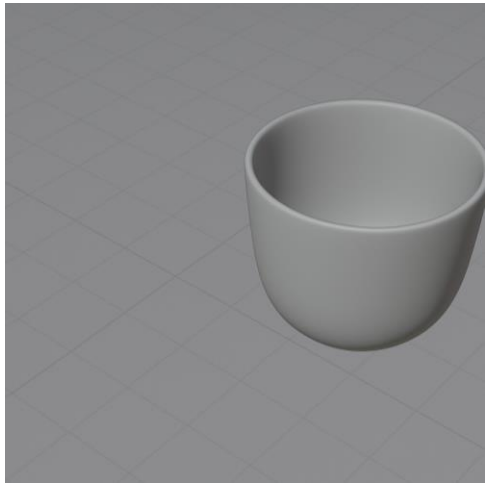
? Tradeoff: Consistency vs. Efficiency

Key idea: Ground pixel generation onto 3D, physical representations.



+





Pixel Generation:
Realism, Diversity

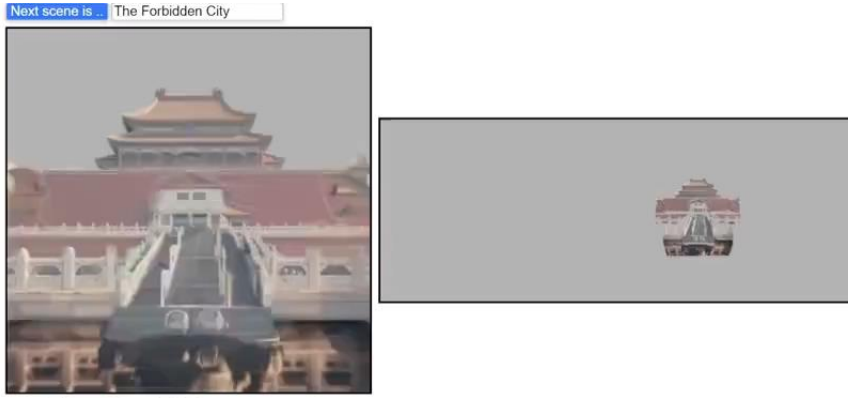
Physical Representation:
Precise Action, Efficient Rendering

Physics-Grounded World Models

Generation

Image

Real-time control



Static 3D World

Interaction

Image

Action

Dynamic 3D World

Evaluation

3D/4D/Video World Models

Benchmark

Interactive 3D World Generation

Goal: Fast 3D scene generation following real-time user control of where to generate what contents.

User Control

Input Image

First-Person View

Bird-Eye View of the World

Yu* et.al., WonderWorld, CVPR'25 (Highlight)

3D World Generation via Grounding

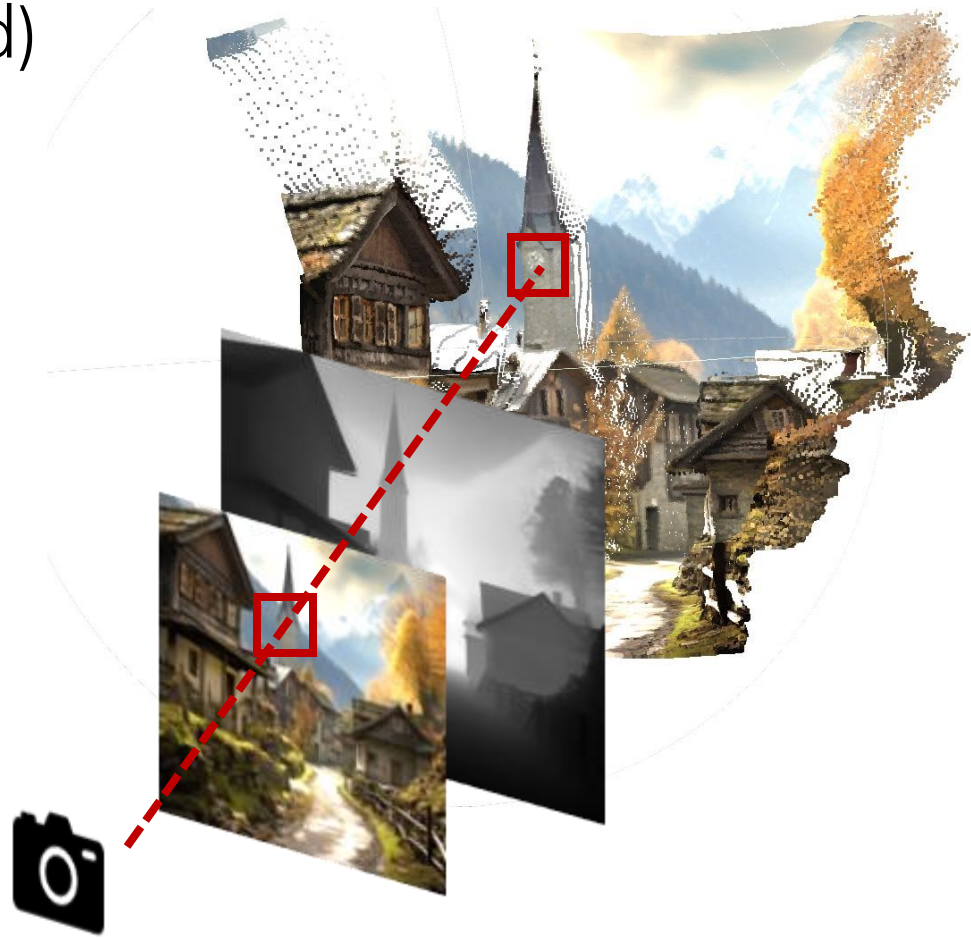
WonderJourney [Yu2024]

RealmDreamer [Shiriram2025]

3D World Generation via Grounding

Use a 3D representation (e.g., point cloud)

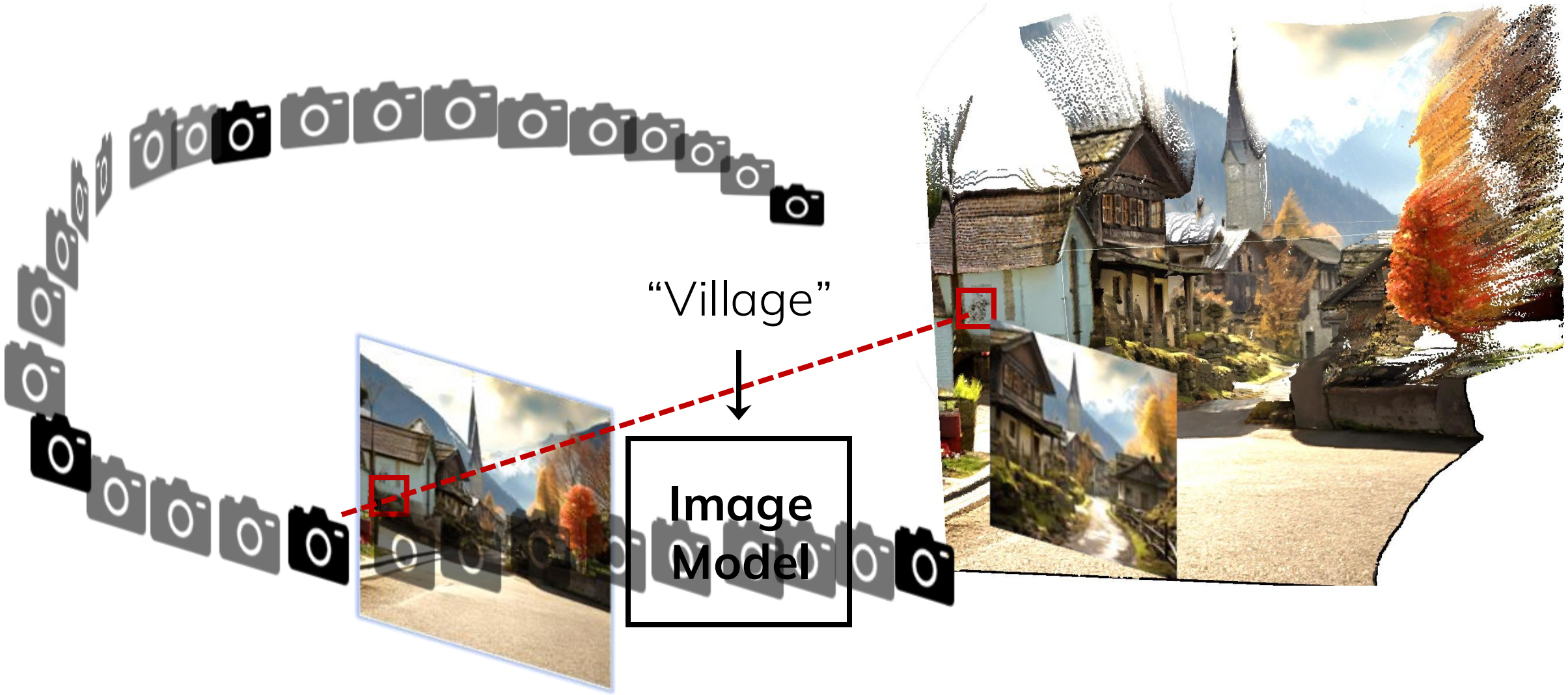
Simple Grounding by Unprojection



Unprojection

Yu et.al., WonderJourney, CVPR'24

3D World Generation via Grounding



Yu et.al., WonderJourney, CVPR'24

Challenge: Too Slow To Be Interactive

RealmDreamer [Shiriram 2025]	**WonderJourney** [Yu 2024]	**WonderWorld**
Hours	749.5s	**9.5s**

Time Cost to Generate a Scene

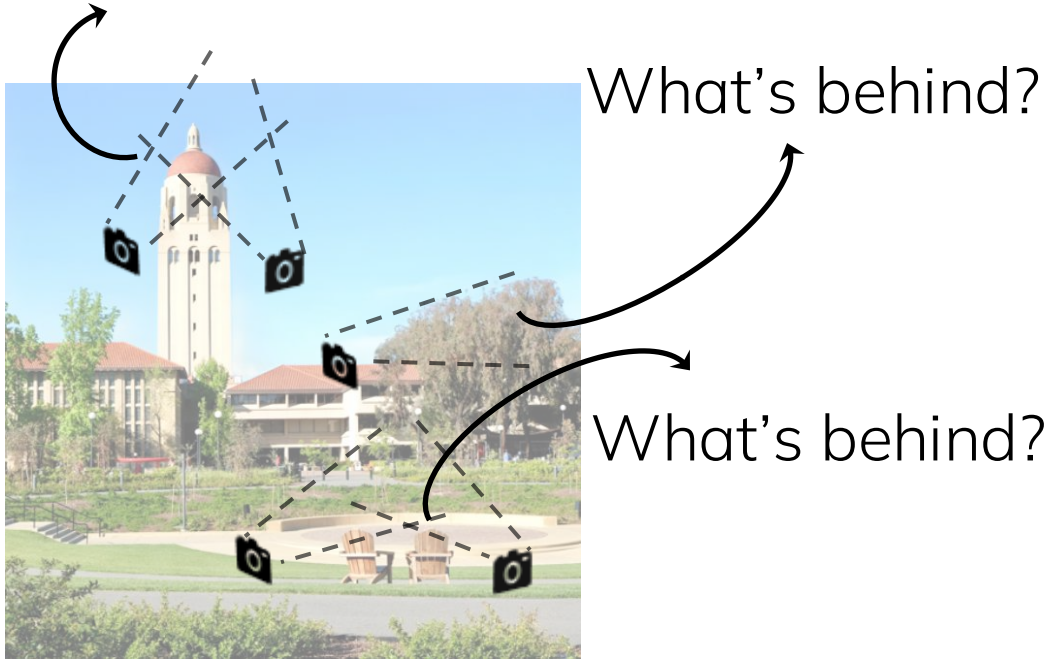
- ❌ Many views to generate
- ❌ Slow 3D optimization (minutes ~ hours)

Fast Layered Gaussian Surfels (FLAGS): Seconds per scene!

- One view
- Fast optimization (<1s)

Challenge: Too Slow To Be Interactive

- Needs to generate many views



Input Image

...

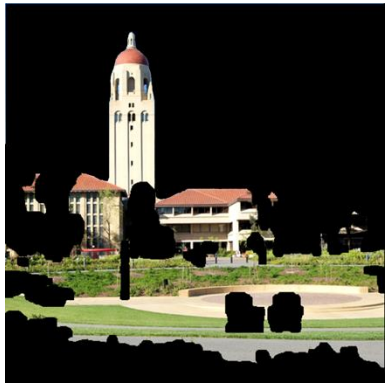
Generated Views

Fast Layered Gaussian Surfels (FLAGS)

Core idea 1: Find occluded regions and complete.

Input Image

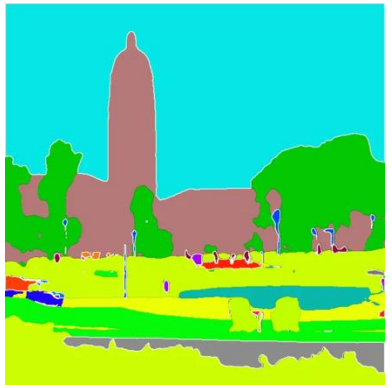
Foreground



Background

Sky

Depth Edge

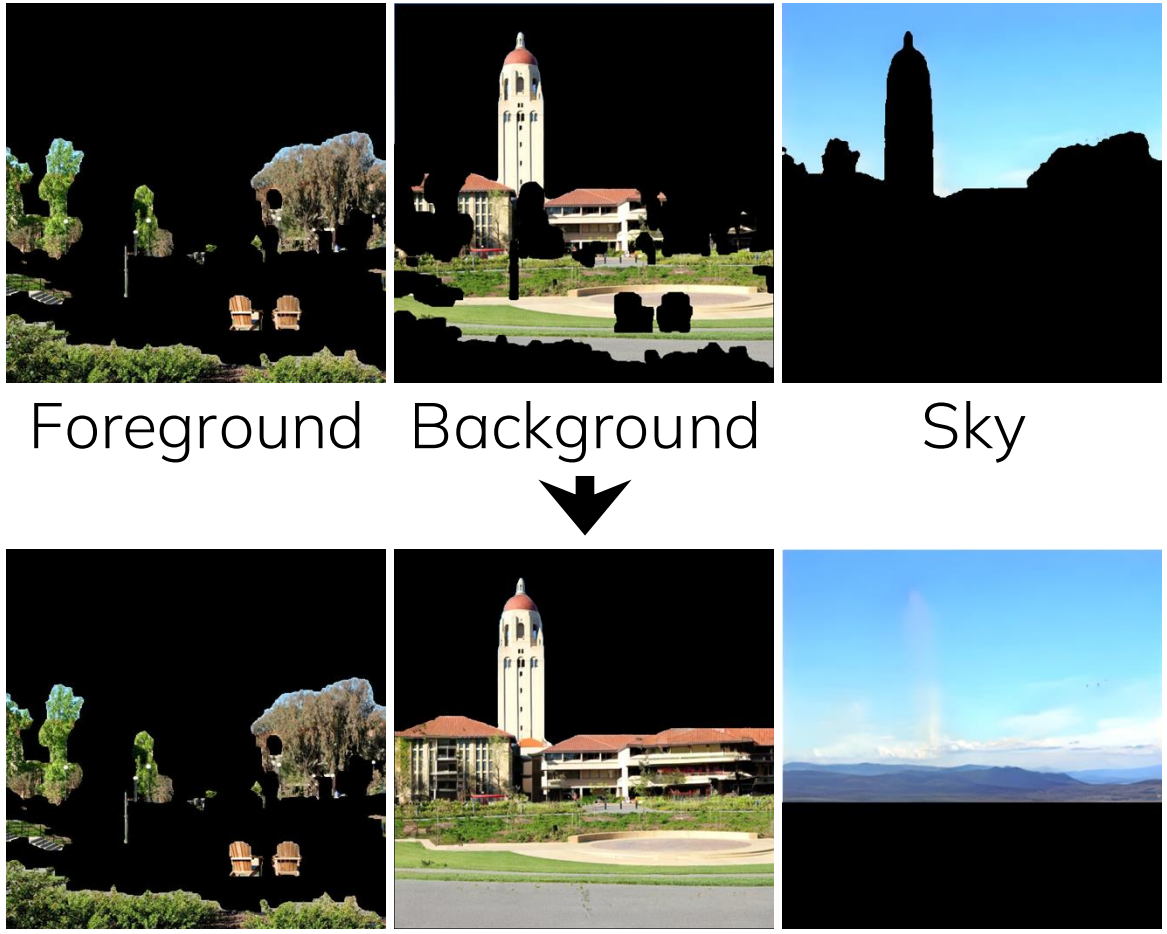


Segment

Fast Layered Gaussian Surfels (FLAGS)

Core idea 1: Find occluded regions and complete.

Input Image



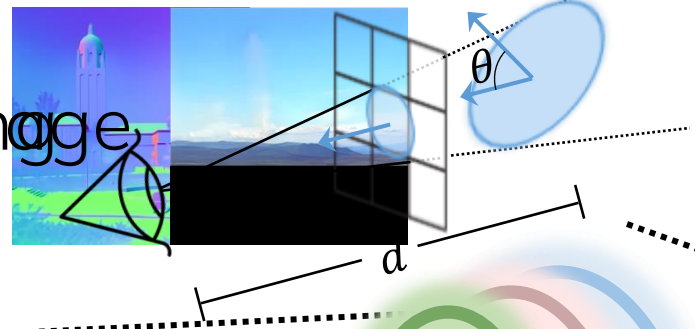
Foreground Background Sky

Completed Layer Images

3D Scene

Fast Layered Gaussian Surfels (FLAGS)

Core idea 2: Design surfels to leverage pixel geometry for initialization.



RGB

XYZ

Radius

Orient.

Opacity

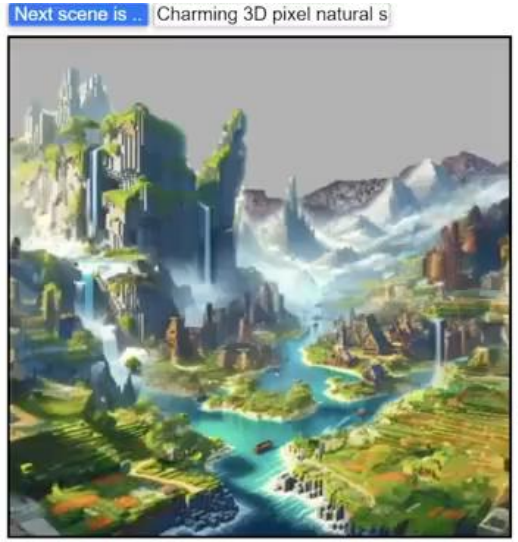
FLAGS rendering a pixel

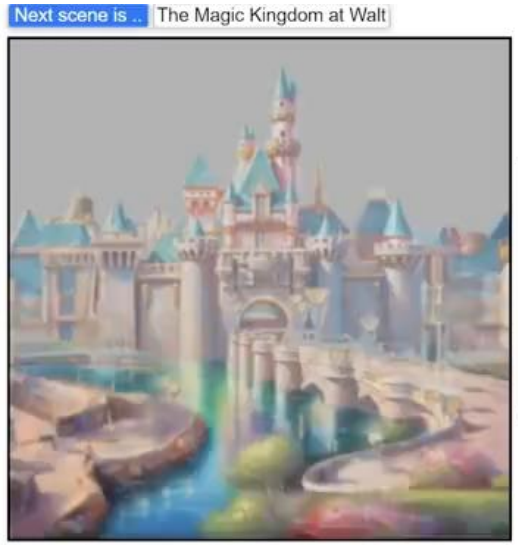
3D Scene

10-Times Slower

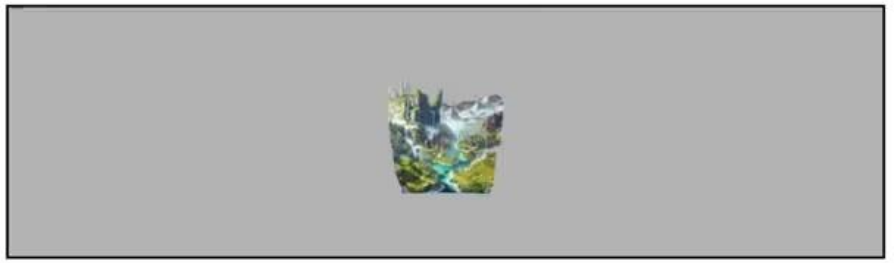
Interactive Generation Process

Input Image



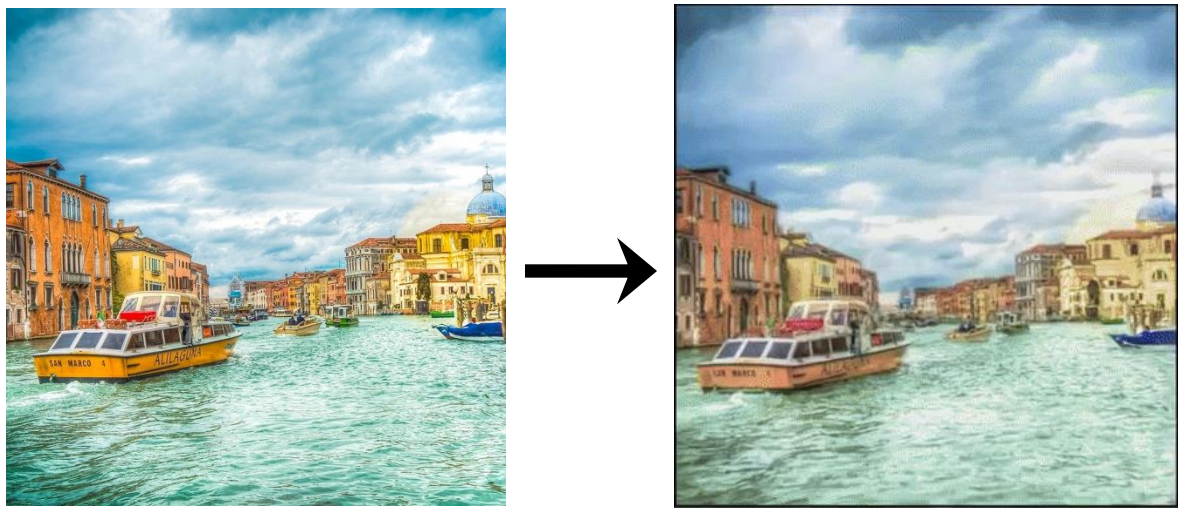


First-Person View

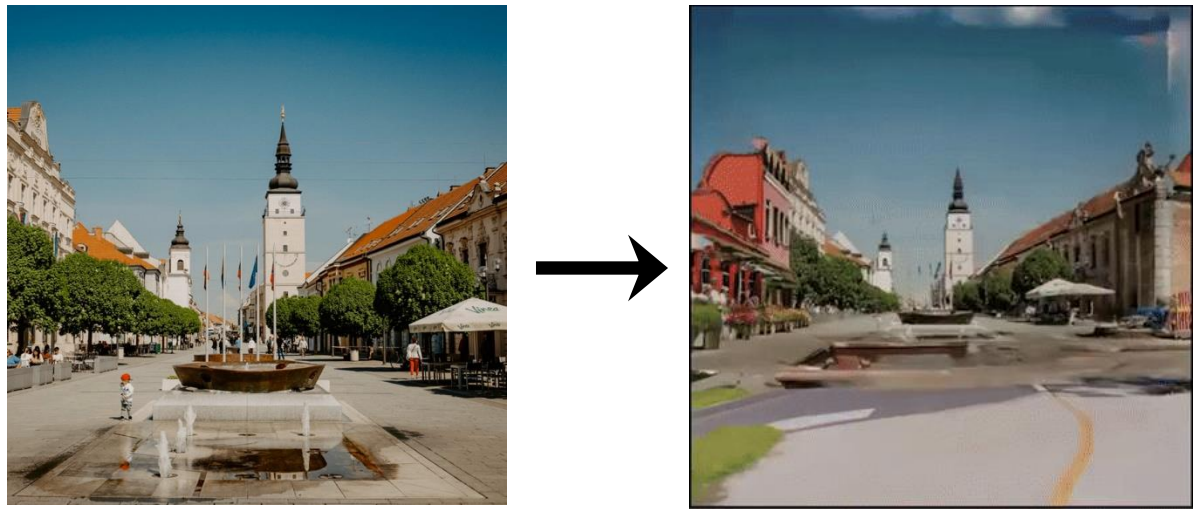


Bird-Eye View

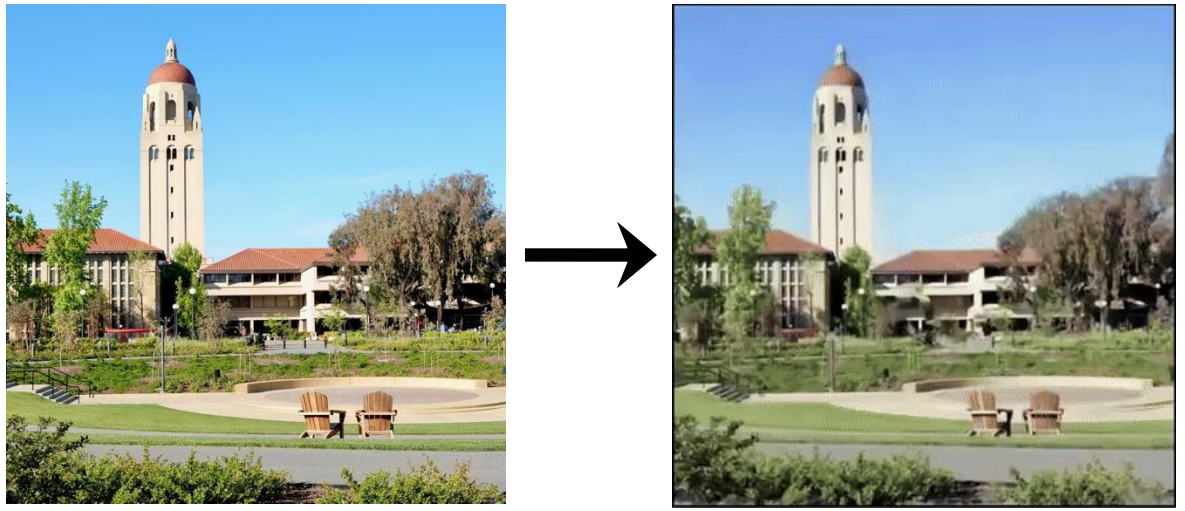
Worlds Generated by Different Users



Panorama Layout



Sector Layout



Winding Layout

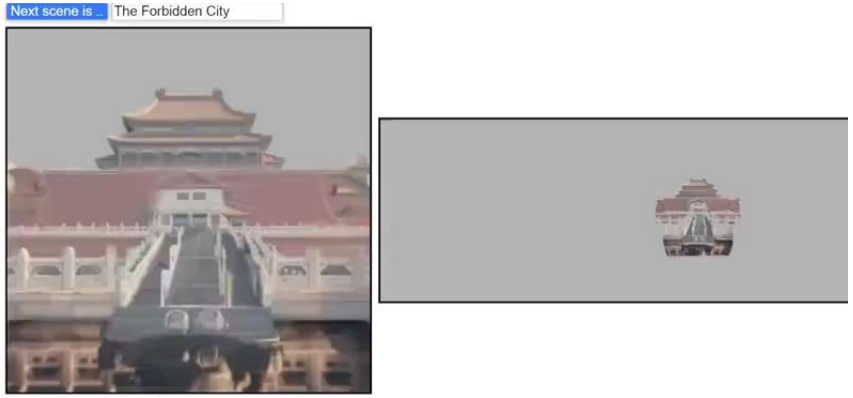
Straight Layout

Physics-Grounded World Models

Generation

Image

Real-time control



Static 3D World

Interaction

Image

Action

Dynamic 3D World

Evaluation

3D/4D/Video World Models

Benchmark

Physics-Grounded World Models

Generation

Image

Real-time control

Static 3D World

Interaction

Image

Action

Dynamic 3D World

Evaluation

3D/4D/Video World Models

Benchmark

Generating Dynamic Scenes under Actions

Li*, Yu*, et.al., WonderPlay, Arxiv'25

Generating Dynamic Scenes under Actions

Li*, Yu*, et.al., WonderPlay, Arxiv'25

Generating Dynamic Scenes under Actions

Li*, Yu*, et.al., WonderPlay, Arxiv'25

Generating Dynamic Scenes under Actions

Li*, Yu*, et.al., WonderPlay, Arxiv'25

Generating Dynamic Scenes under Actions

Li*, Yu*, et.al., WonderPlay, Arxiv'25

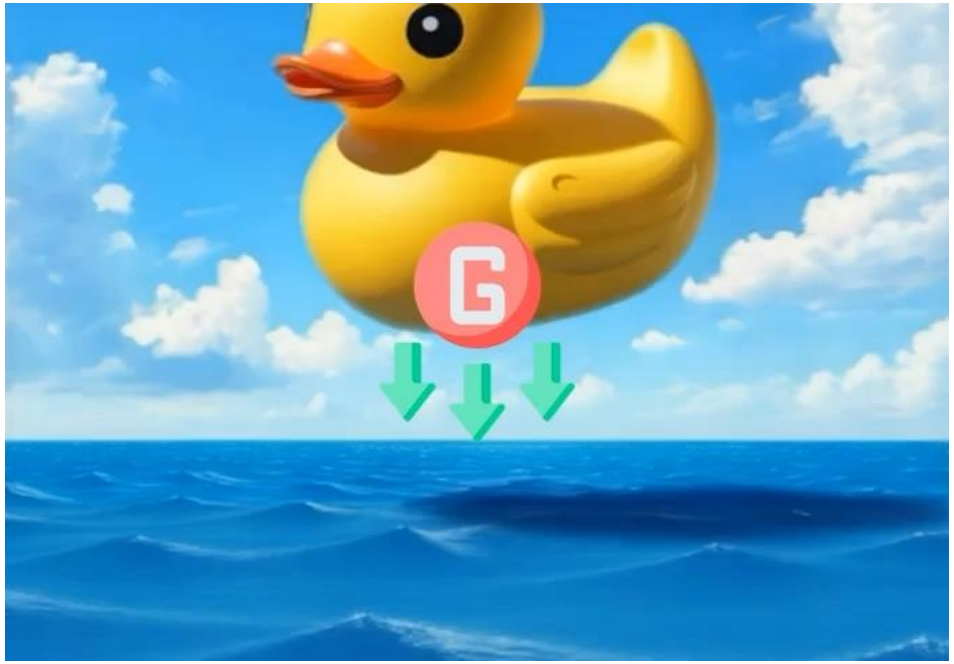
Generating Dynamic Scenes under Actions

Li*, Yu*, et.al., WonderPlay, Arxiv'25

Generating Dynamic Scenes under Actions

Goal: Predicting physical dynamics of generated scenes under applied 3D actions.

Input Image

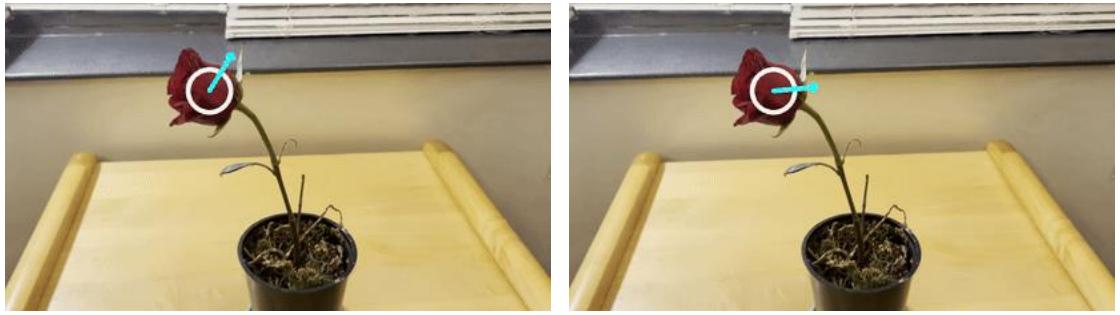


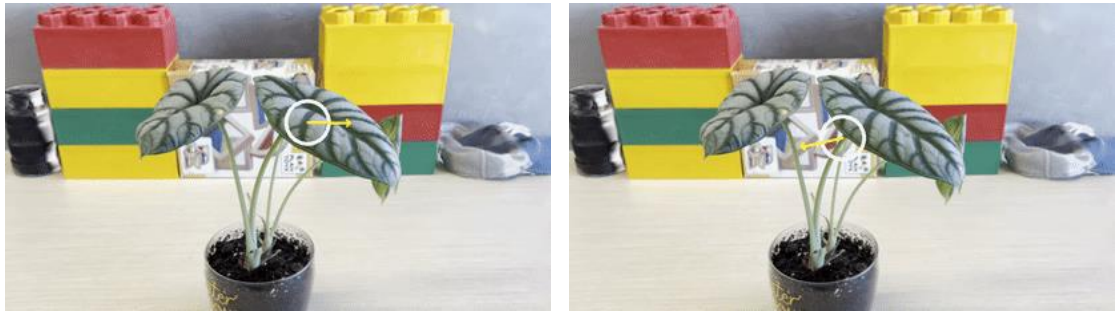
Interactive 3D Scene

Li*, Yu*, et.al., WonderPlay, Arxiv'25

Prior Physics-Grounded World Models

Relying on physics simulation to generate future dynamics.





3DGS + Material

Physics Simulator

Action

Next state $\boldsymbol{x}_{t+1}$

Render

Dynamic Scene

Zhang, Yu, et.al., PhysDreamer, ECCV'24 (Oral)

Challenge: Multi-Physics Simulation is Hard

- **Inaccuracy**: Multi-physics simulation is not accurate, even with perfect materials.
- Hard to get **full** physical states from a single image.

WonderPlay (Ours)

PhysDreamer

PhysGen

Liu, et.al., PhysGen, ECCV'24

Video model to improve multi-physics simulation?

Challenge: Video models cannot take **actions** as input.

{Action, Video} pairs for post-training? Too few.

Core idea: Hybrid model, where simulator predicts action-conditioned dynamics, and video model refines dynamics in-the-loop.

WonderPlay: Hybrid Generative Simulator

Input Image

FLAGS

3D Scene

Hybrid Generative Simulator

Physics Simulator

Video Generator

Refined Simulation

Generated Video

Bimodal Control to Condition Video Gen

Simulated RGB

Structured Noise

Simulated Motion

Video Generator

Refined Simulation

Generated Video

Photometric Loss

Diverse Physics in Interactions

Different Actions

Physics-Grounded World Models

Generation

Image

Real-time control

Static 3D World

Interaction

Image

Action

Dynamic 3D World

Evaluation

3D/4D/Video World Models

Benchmark

Physics-Grounded World Models

Generation

Image

Real-time control

Static 3D World

Interaction

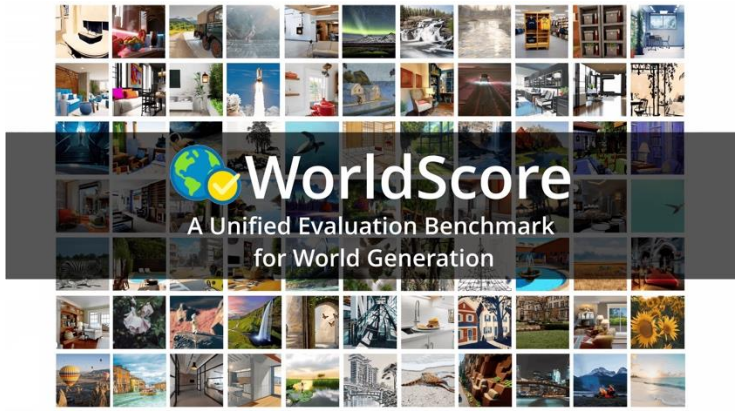
Image

Action

Dynamic 3D World

Evaluation

3D/4D/Video World Models



Benchmark

Existing Eval: Small Scenes, VideoGen Only

✓ Video Models

✗ 3D Models

✗ 4D Models

Generating & Evaluating Small Scenes Only

Can’t Eval 3D/4D

Problem: Semantic prompt only. No **spatial prompts.**

Existing Eval: Small Scenes, VideoGen Only

Existing benchmarks: Semantic prompt ("**what**"), no spatial prompt ("**where**").

- A large world requires **spatial prompts** of each scene.
- 3D/4D models need them as input.

"A tranquil tableau of an ancient... The scene captures a sense of ..."

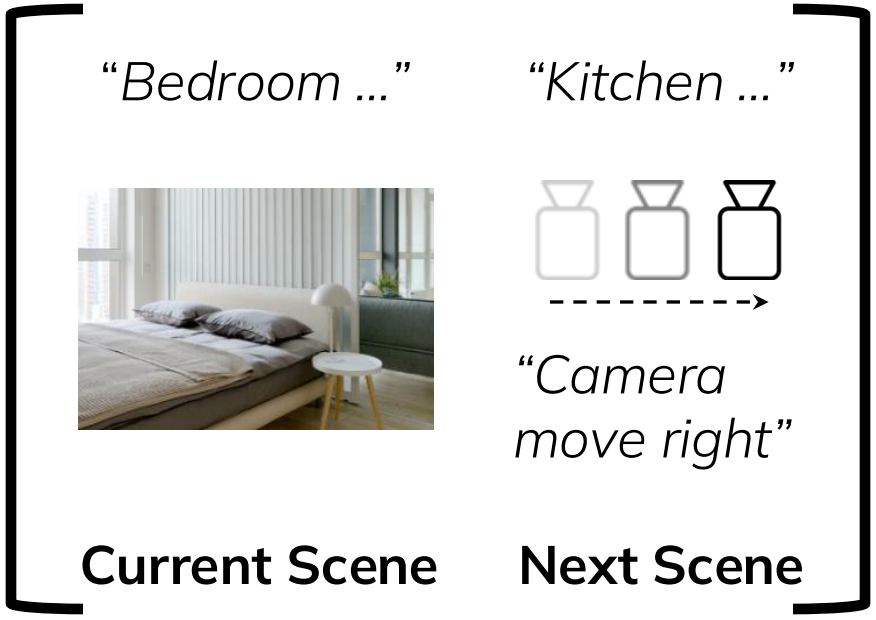
Semantic Prompt

A World of Multiple Scenes

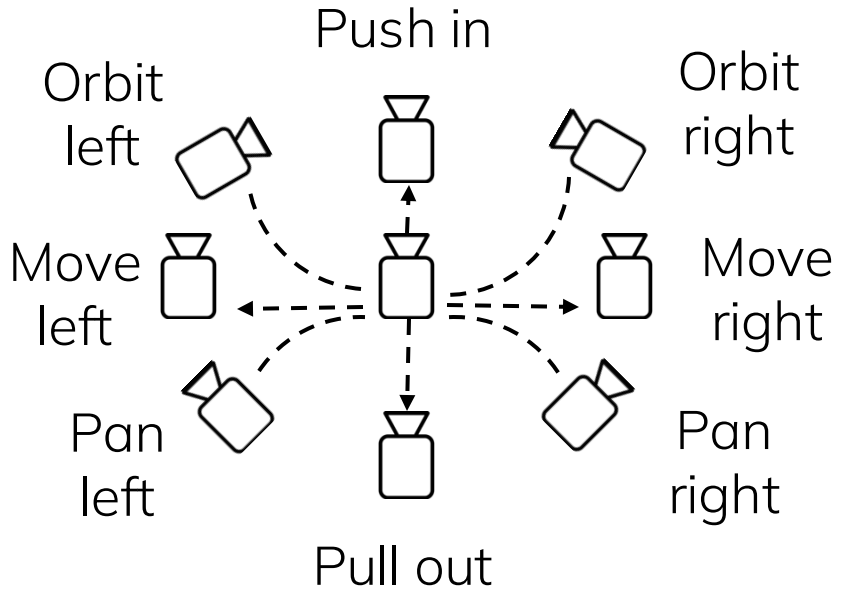
Challenge 1: Unified World Specification

Problem: Prompt 3D/4D/video models in a unified way.

Key idea: Decompose world generation as a sequence of **next-scene generation** tasks, each described by **semantic and spatial prompts**.



A Generated Example



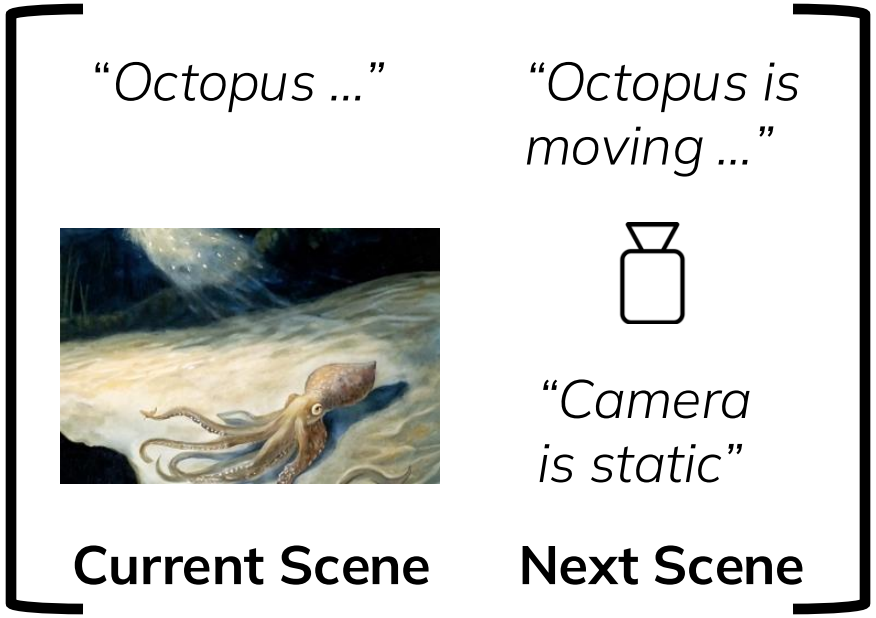
Spatial Prompt Library

Challenge 2: Camera Motion vs. Scene Motion

Problem: Camera motion (to eval physical consistency) and scene motion (to eval dynamics generation) are entangled.

Key idea: Evaluate dynamics with fixed cameras; evaluate generation with static scenes.

Entangled Motion



Generated Example

Diverse Datasets for Static & Dynamic Worlds

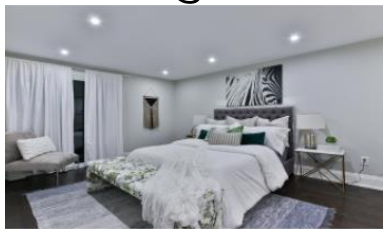
Scene image: Categorization + Extensive data sources + Filtering + VLM captioning

Next-scene prompt: LLM to generate diverse semantic prompts.

Indoor:

Dining

Living



Passageway

Public

Workspace

Outdoor:

City

Suburb

Aquatic

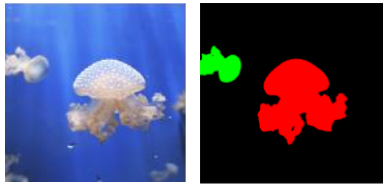
Terrestrial

Verdant

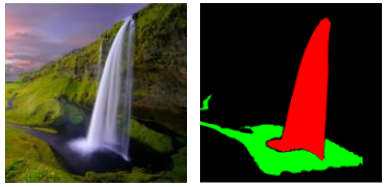
Motion:

Articulated

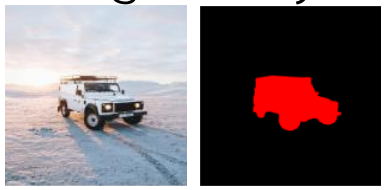
Deformable



Fluid



Rigid body



Multi-motion

WorldScore Metrics

Controllability

- Camera Controllability
- Object Controllability
- Content Alignment

Quality

- 3D Consistency
- Photometric Consistency
- Style Consistency
- Subjective Quality

Dynamics

- Motion Accuracy
- Motion Magnitude
- Motion Smoothness

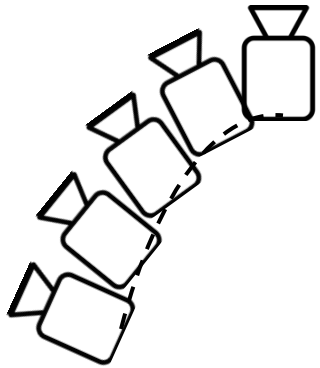
WorldScore Metrics: Controllability

Controllability
- Camera Controllability
- Object Controllability
- Content Alignment

Quality

Dynamics

"Camera pans left"



Higher Score

Lower Score

WorldScore Metrics: Controllability

Controllability | Quality | Dynamics

- Camera Controllability
- Object Controllability
- Content Alignment

- Pancakes
- Orange Juice

→

Higher Score

Lower Score

WorldScore Metrics: Controllability

Controllability | Quality | Dynamics

- Camera Controllability
- Object Controllability
- Content Alignment

"Lockers, trophy, courtyard, fountain, benches ..."

Higher Score

Lower Score

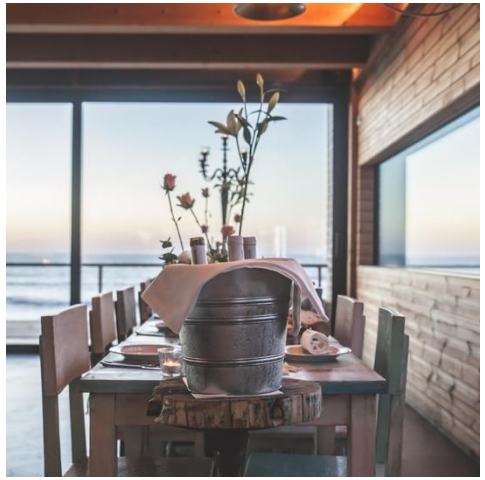
WorldScore Metrics: Quality

Controllability

Quality

- 3D Consistency
- Photometric Consistency
- Style Consistency
- Subjective Quality

Dynamics



"Seaside dining ..."

Higher Score

Lower Score

WorldScore Metrics: Quality

Controllability

Quality

- 3D Consistency
- Photometric Consistency
- Style Consistency
- Subjective Quality

Dynamics

"Mountain range, clouds caress the peaks ..."

Higher Score

Lower Score

WorldScore Metrics: Quality

Controllability

Quality

- 3D Consistency
- Photometric Consistency
- Style Consistency
- Subjective Quality

Dynamics

"Urban Street View ..."

Higher Score

Lower Score

WorldScore Metrics: Quality

Controllability

Quality

- 3D Consistency
- Photometric Consistency
- Style Consistency
- Subjective Quality

Dynamics

"A bright modern kitchen ..."

Higher Score

Lower Score

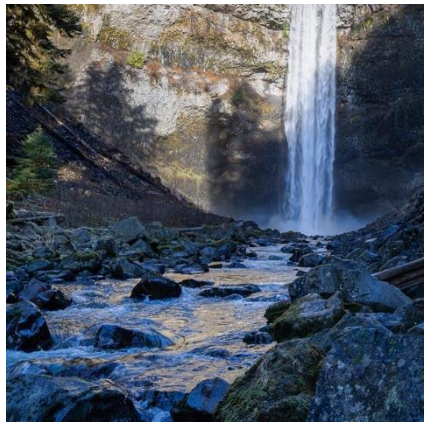
WorldScore Metrics: Dynamics

Controllability

Quality

Dynamics

- Motion Accuracy
- Motion Magnitude
- Motion Smoothness



Waterfall plunges, the stream flows ...

Higher Score

Lower Score

WorldScore Metrics: Dynamics

Controllability

Quality

Dynamics

- Motion Accuracy
- Motion Magnitude
- Motion Smoothness

“Octopus glide as waves lap …”

Higher Score

Lower Score

WorldScore Metrics: Dynamics

Controllability

Quality

Dynamics

- Motion Accuracy
- Motion Magnitude
- Motion Smoothness

"Panda climbs ..."

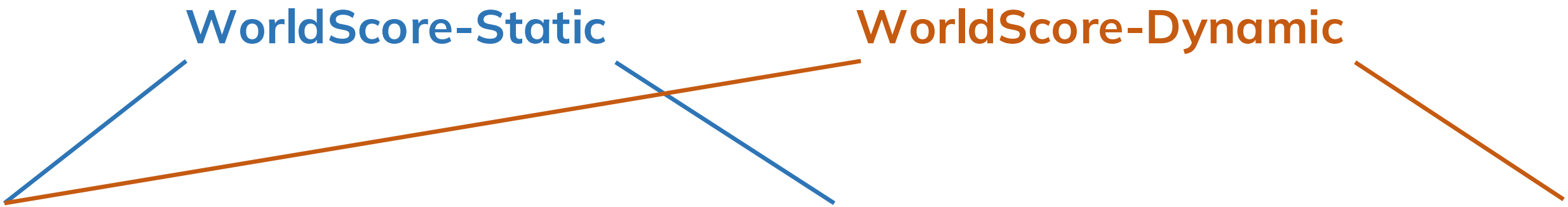
Higher Score

Lower Score

WorldScore Metrics

	# Examples	Multi-Scene	Unified	Camera Ctrl.	3D Consist.
VBench	800	❌	❌	❌	❌
EvalCrafter	700	❌	❌	❌	❌
T2V-CompBench	700	❌	❌	❌	❌
TC-Bench	150	❌	❌	❌	❌
WorldModel Bench	350	❌	❌	❌	❌
WorldScore (Ours)	**3000**	✅	✅	✅	✅

WorldScore Metrics



Controllability
- Camera Controllability
- Object Controllability
- Content Alignment

Quality
- 3D Consistency
- Photometric Consistency
- Style Consistency
- Subjective Quality

Dynamics
- Motion Accuracy
- Motion Magnitude
- Motion Smoothness

We evaluate 19 models including video/3D/4D models.

Evaluation: Takeaway Message 1

Message 1: 3D models excel in static world generation.

	CogVideoX-I2V	Vchitect-2.0	LucidDreamer	WonderWorld
WorldScore-Static	62.15	42.28	<u>70.40</u>	**72.69**

3D Model Video Model

Evaluation: Takeaway Message 2

Message 2: The best open-source video models are as good as closed-source video models.

	Gen-3	Hailuo	CogVideoX-I2V
WorldScore-Static	60.71	57.55	**62.15**
WorldScore-Dynamic	57.58	56.36	**59.12**
Open-Source?	No	No	**Yes**

Evaluation: Takeaway Message 3

Message 3: Video models are weak in generating larger worlds.

	Gen-3	DynamiCrafter	VideoCrafter1-I2V
Worlds of 2 Scenes	64.71	56.12	58.71
Worlds of 4 Scenes	46.94	37.01	19.83

Low Camera Controllability

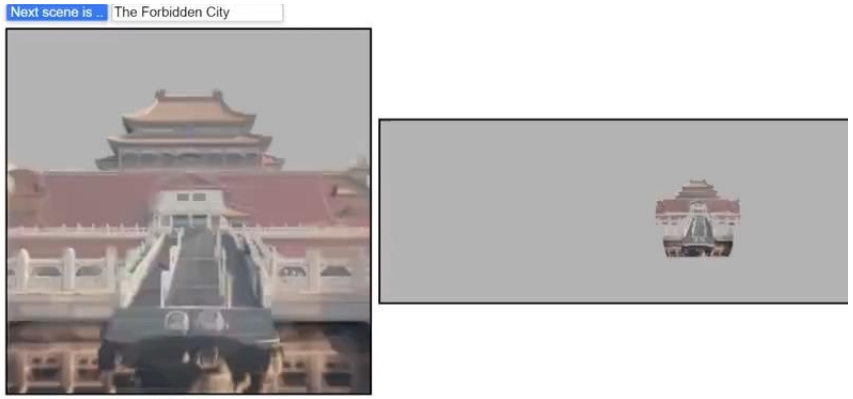
Quality Degradation over Time

Physics-Grounded World Models

Generation

Image

Real-time control



Static 3D World

Interaction

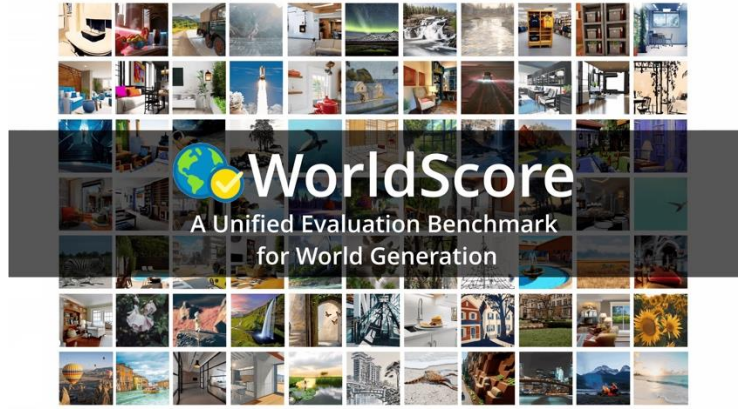
Image

Action

Dynamic 3D World

Evaluation

3D/4D/Video World Models



Benchmark

Wonderful Collaborators!