如何构建既真实又可交互的生成式世界

—— 从视频生成到显式3D：世界模型的两条探索路线

孙阳天

CVMI Lab @ HKU

https://sunyangtian.github.io/

Introduction: Toward Interactive World Models

What should a world model deliver?

DreamX-World · click to play

Visual realism

Convincing appearance and motion

WorldMem · revisit video

Visual consistency

Same world across time and viewpoints

DreamDojo · live teleoperation

Controllability

Actions determine what happens next

Matrix-Game 3.0 · click to play

Real-time interaction

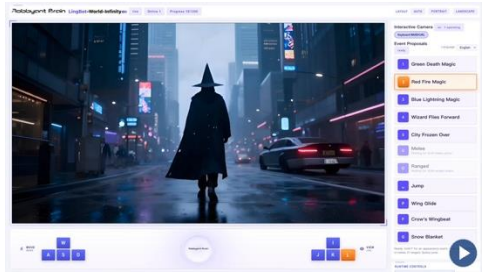
Act → observe → act
Low-latency simulation

Realistic observations. Persistent worlds. Responsive actions.

Video-first: Recent World Models

Representative interactive world models

LingBot-World 2.0



Robbyant · Jul 2026

Long-horizon exploration
Richer actions and
shared worlds

Matrix-Game 3.0

Skywork AI · 2026

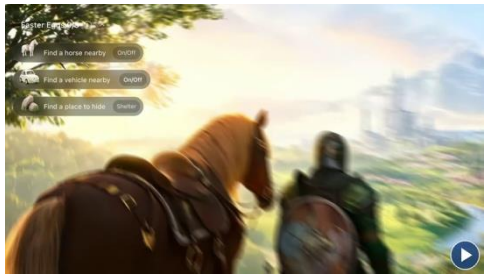
Camera-aware memory
Streaming keyboard /
mouse control

DreamX-World 1.0

DreamX Team · Jun 2026

Camera control
Scene memory
Composable events

Happy Oyster



Official project · 2026

Directing + Wandering
Text / voice / image
steering

Official demos · click each video to play; not a shared benchmark

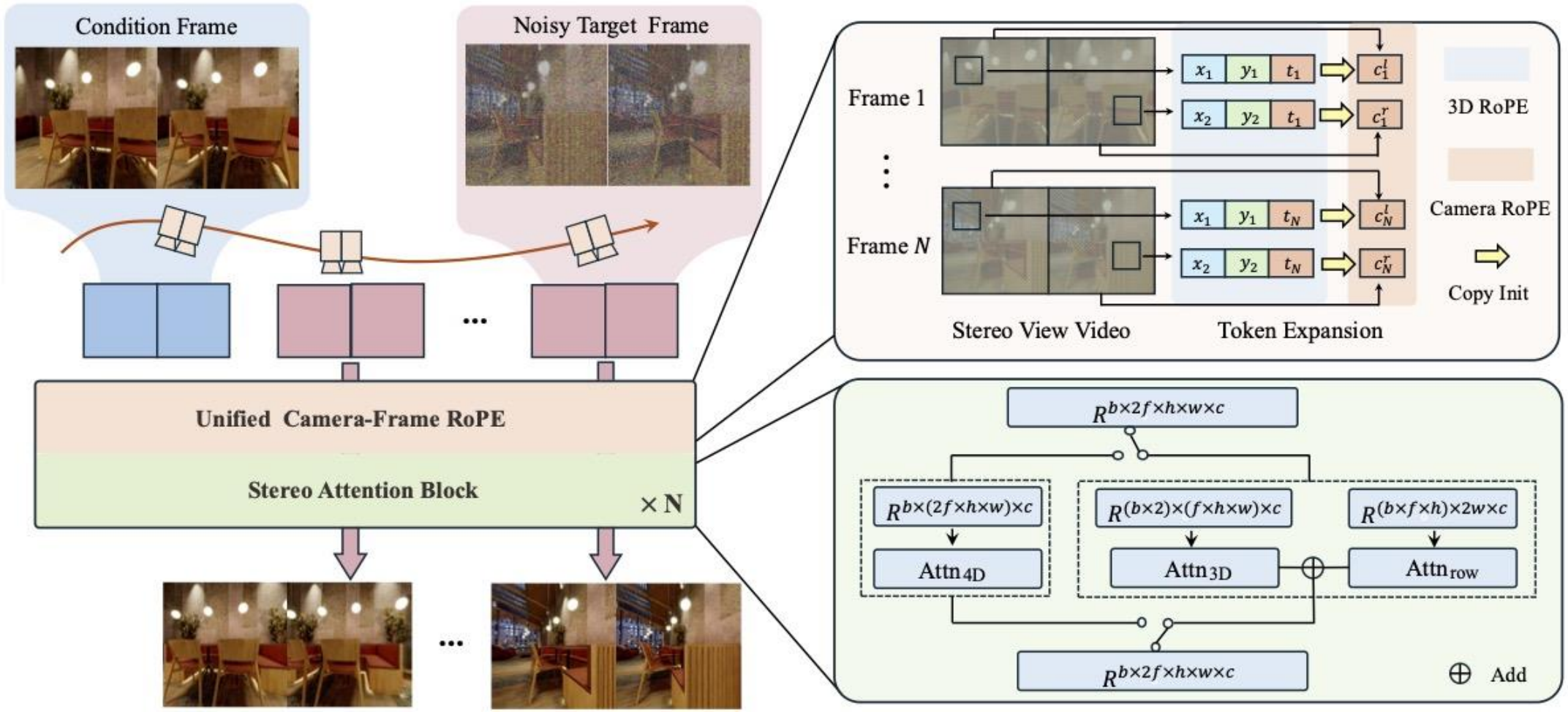
Video-first: Stereo World Model

- **From monocular to stereo**

World Model Modality	Monocular, RGB	Monocular, RGBD
Geometry	❌	Relative, Limited Range
Same Domain with Pre-Trained Model	✅	❌
VR Display	Rely on External Depth Estimation and Inpainting Model	

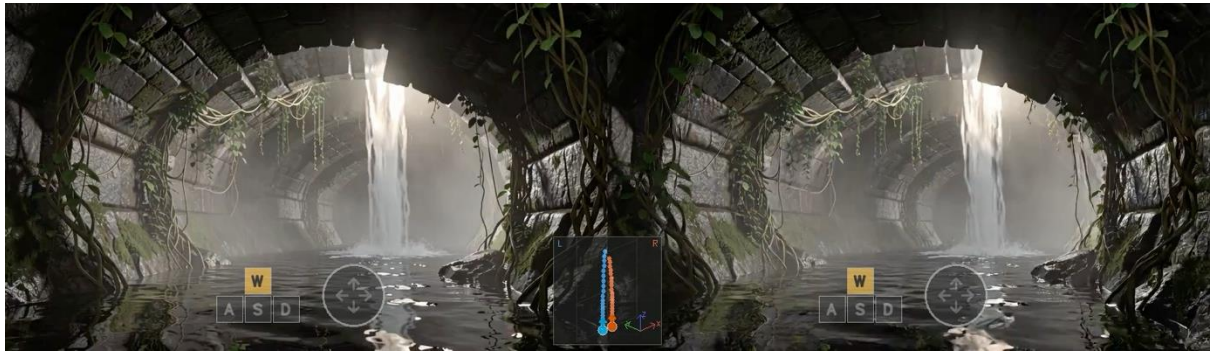
Video-first: Stereo World Model

- **Stereo World: Pipeline**



Video-first: Stereo World Model

- **Fixed-baseline Stereo-Video**

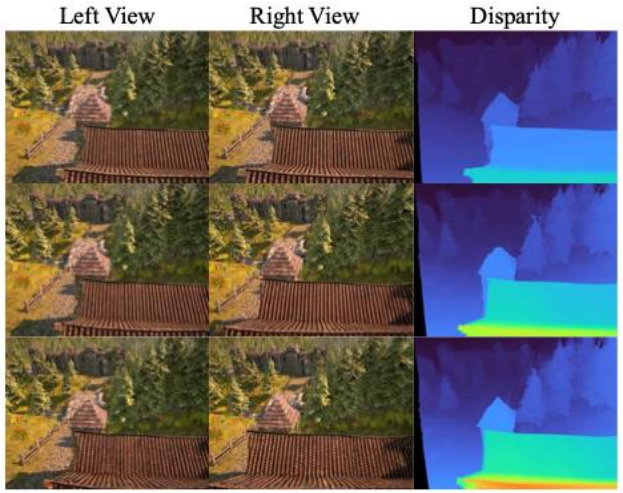


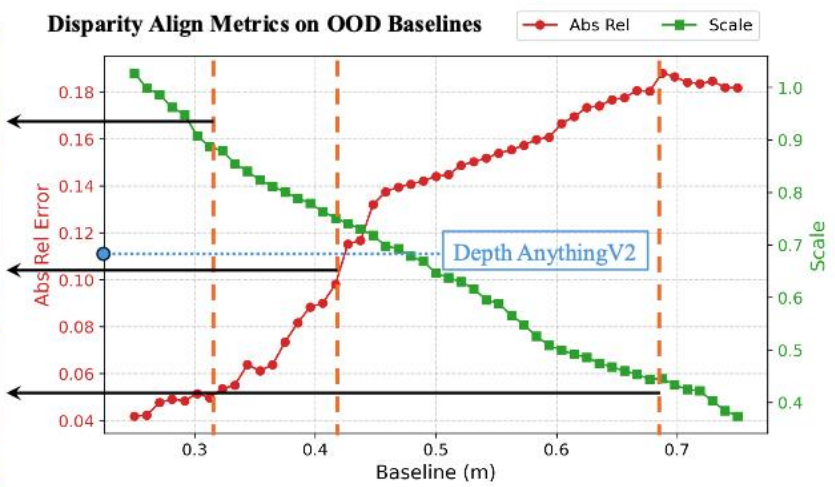
Video-first: Stereo World Model

Beyond simply adding a view

Method	Modality	Visual Quality				Camera Accuarcy		View Synchronization			FPS↑
		FID↓	FVD↓	CLIP-T↑	CLIP-F↑	RotErr↓	TransErr↓	Mat. Pix.(K)↑	FVD-V↓	CLIP-V↑	
Voyager [26]	RGBD	226.97	170.37	24.85	97.03	1.34	0.25	4.26	55.45	91.41	0.03
DeepVerse [10]	RGBD	191.32	176.72	24.59	97.31	1.51	0.16	4.48	33.50	93.86	0.35
Aether [52]	RGBD	185.72	152.97	24.93	97.14	1.50	0.13	4.35	42.07	93.71	0.11
SEVA [81]	RGB	195.70	170.92	24.77	98.11	1.09	0.51	4.49	31.10	94.73	0.10
ViewCrafter [75]	RGB	211.89	185.76	25.02	96.15	1.24	0.20	4.49	42.10	93.51	0.13
Ours Monocular	*RGB*	*126.83*	*96.87*	*24.97*	*97.12*	*1.36*	*0.14*	✗	✗	✗	✗
Ours Stereo	RGB	111.36	83.04	25.74	97.55	1.01	0.11	4.56	22.00	97.50	0.49

Better view-consistency compared warp-and-inpaint && Better visual quality and camera control compared monocular setting





StereoWorld handles video sequences with varying stereo baselines,

despite being trained only on fixed-baseline data.

Video-first: Stereo World Model

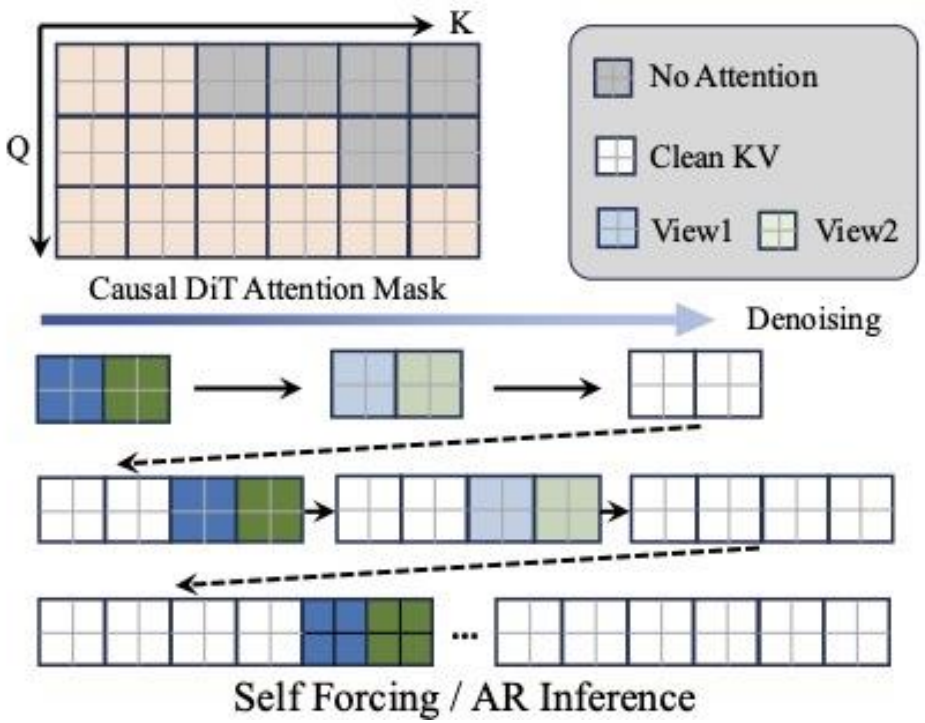
- **From Fixed Stereo-Video to Flexible Stereo-Video**

Video-first: Stereo World Model

- **From Joint Stereo Video Generation to View Inpainting** $p(V_L, V_R \mid C_L, C_R) \rightarrow p(V_R \mid V_L, C_L, C_R)$

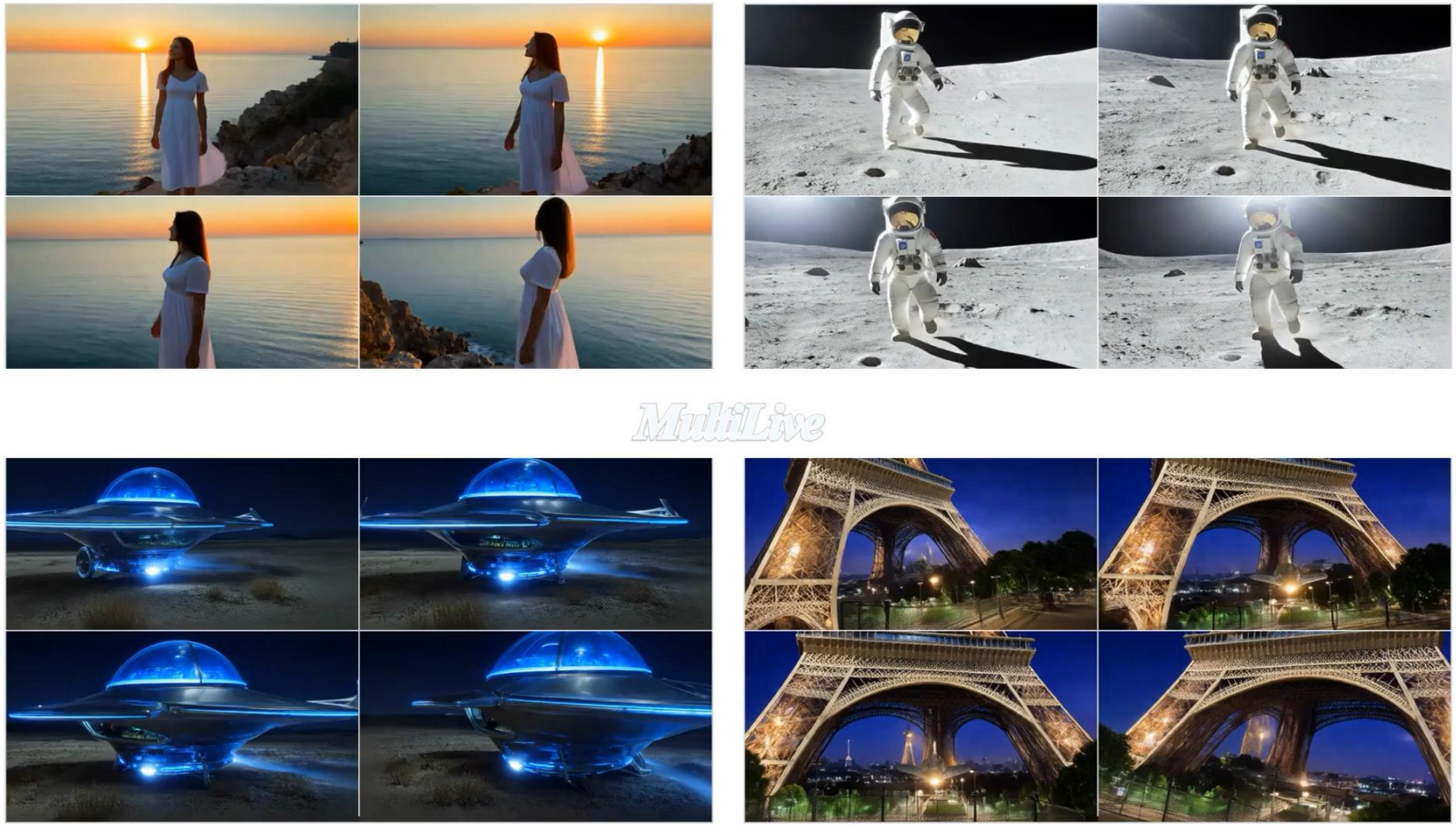
Video-first: Stereo World Model

- **Autoregressive Video Generation**



Video-first: Multi-Live

- **Stereo / Multi-View Video Data is limited, how to solve this ?**

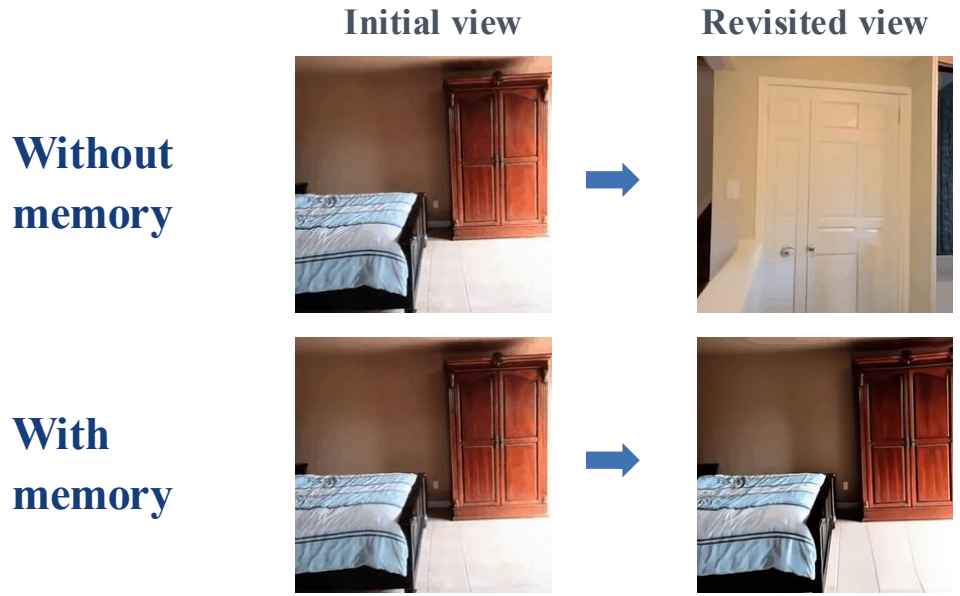


Video-first: Two Remaining Challenges

Memory consistency and complex physical interaction

01 Memory consistency

Does a place stay the same
after leaving and returning?

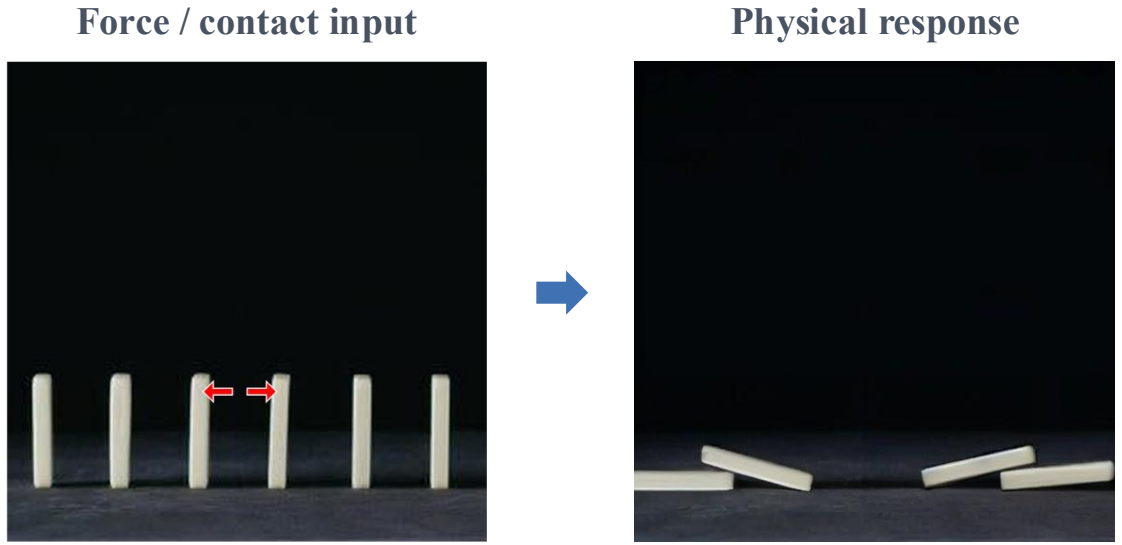


WorldMem (2025) · 360° revisit ablation

Short-term smoothness is not long-term memory.

02 Physics Plausibility

"Push the domino" describes a goal,
not the contact, force or motion.



PhysGen (ECCV 2024) · example of explicit physical control

Beyond prompts and navigation:
continuous actions, contact and feedback.

3D-first: Building Explicit and Interactive Worlds

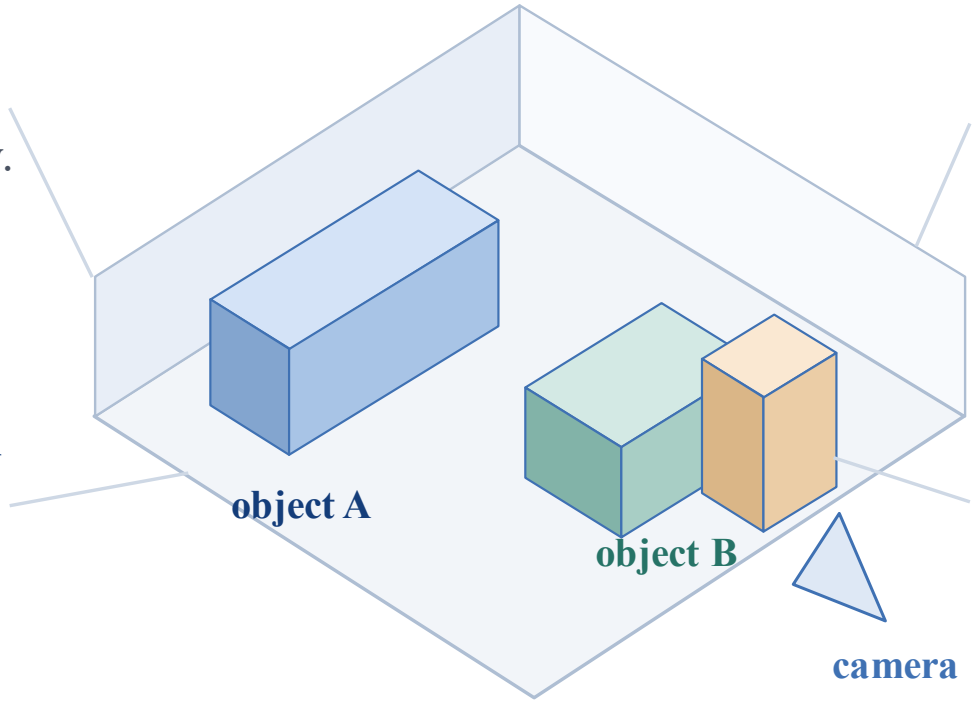
Why Explicit 3D Worlds?

Persistent scene structure

Store and update world state independently of the current view.

Free camera control

Render from a new viewpoint using the same scene.



Independent object control

Change one object's pose without rebuilding the scene.

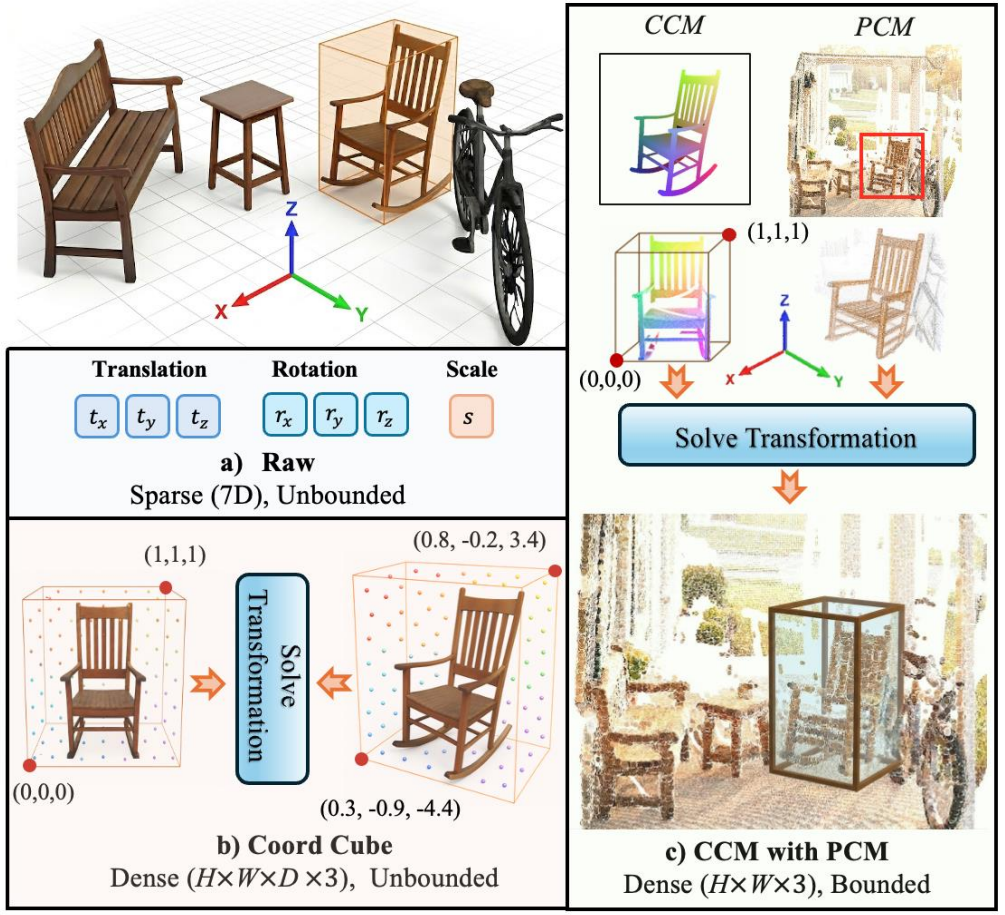
Physics compatibility

Attach colliders and dynamics for simulated interaction.

Objects + transforms + state + physical properties

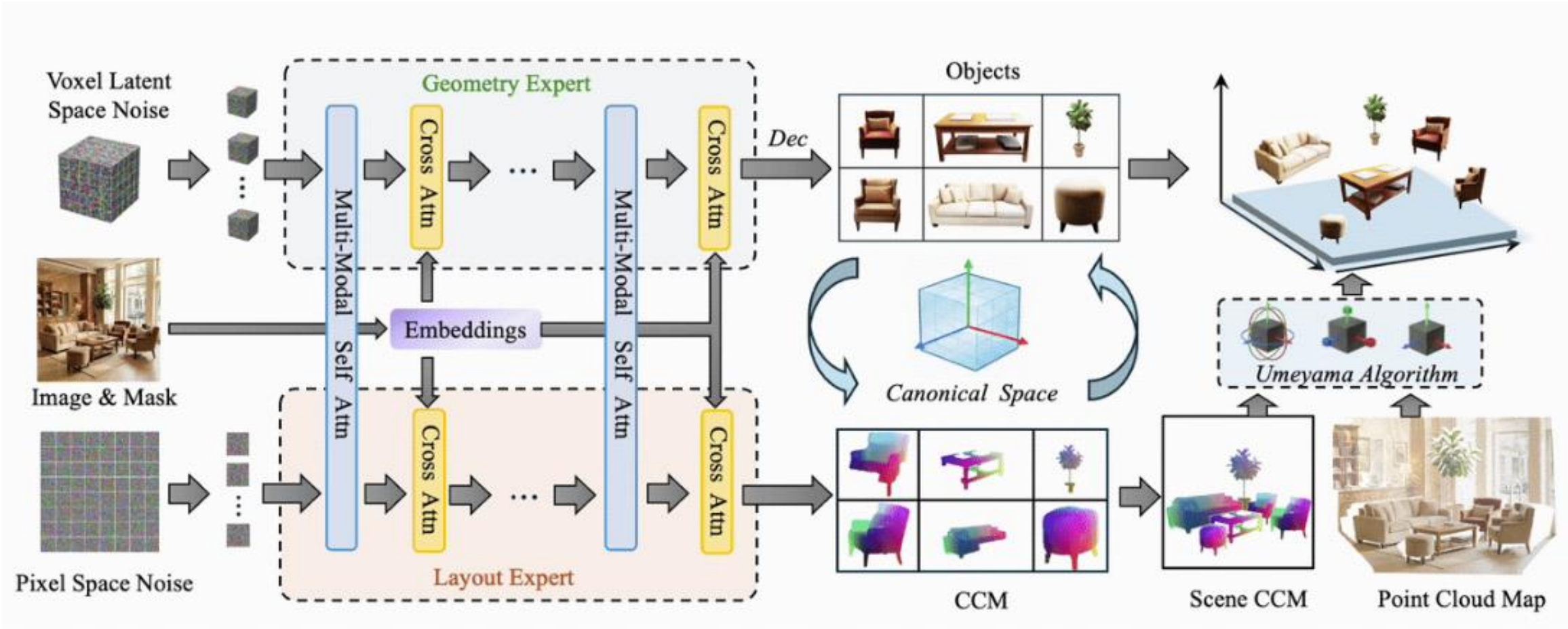
3D-first: Mira-Scene

- **Pixel-Aligned Layouts for Generative 3D Scene Reconstruction**



3D-first: Mira-Scene

- **Pixel-Aligned Layouts for Generative 3D Scene Reconstruction**



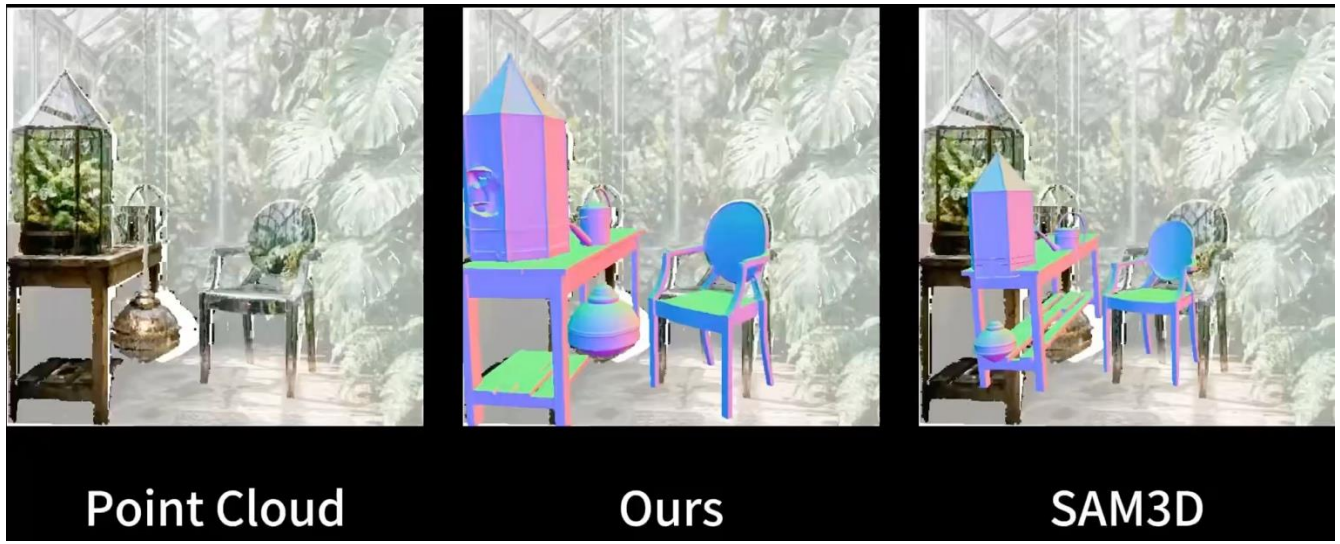
3D-first: Mira-Scene

- **From Single Image to Interactive 3D Scene**

3D-first: Mira-Scene

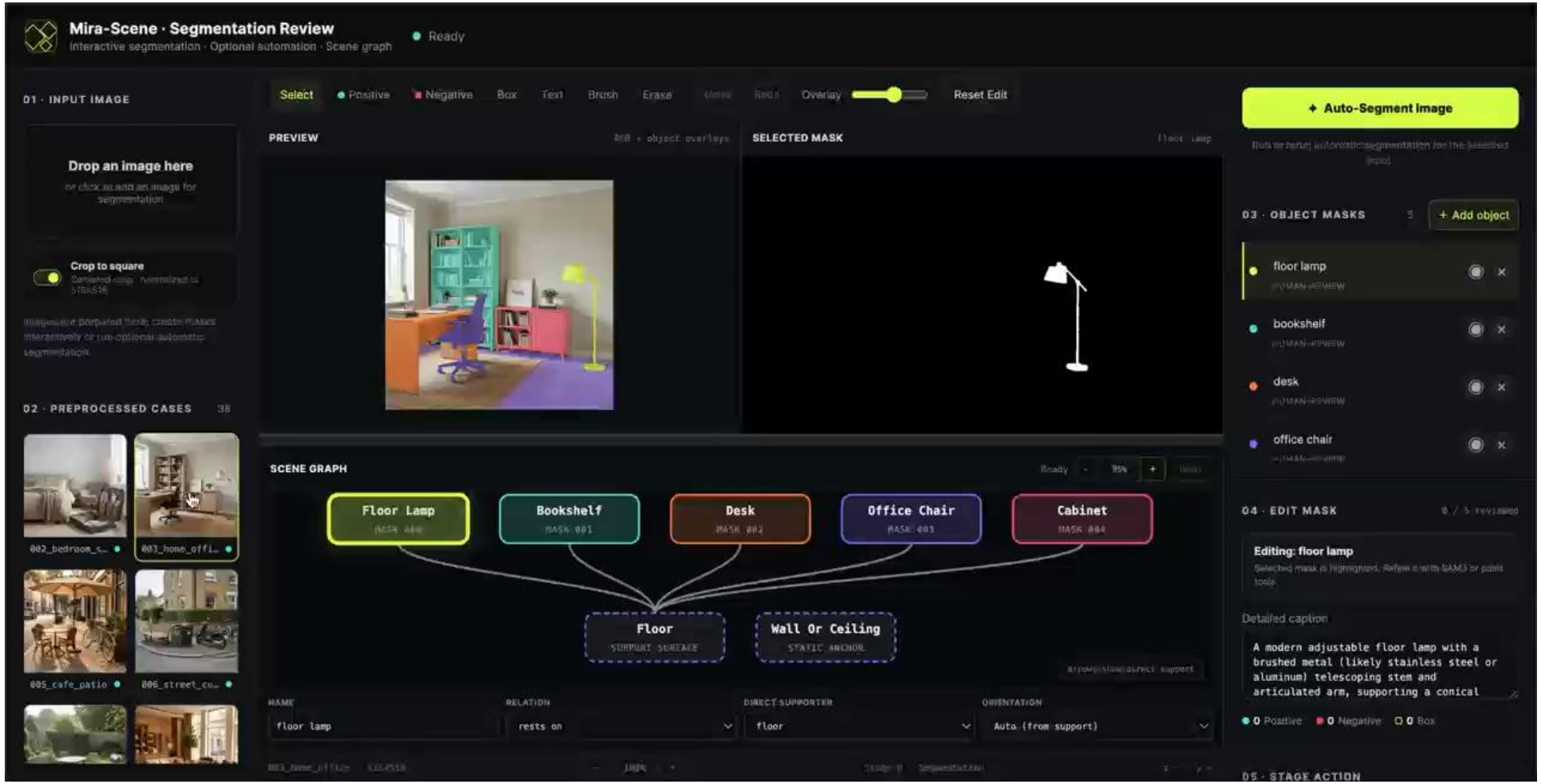
- **From Single Image to Interactive 3D Scene**

Method	BlendSwap							3D-Future Scene						
	CD↓	FS@0.1↑	EMD↓	3D-IoU↑	ICP-Rot↓	2D-IoU↑	ADD-S↓	CD↓	FS@0.1↑	EMD↓	3D-IoU↑	ICP-Rot↓	2D-IoU↑	ADD-S↓
Gen3DSR	0.072	0.582	0.263	0.359	15.54	0.608	0.163	0.067	0.576	0.263	0.502	11.37	0.638	0.107
MIDI	0.036	0.768	0.230	0.229	9.366	0.472	0.449	0.039	0.761	0.249	0.280	14.53	0.387	0.164
SceneGen	0.032	0.794	0.238	0.185	15.08	0.394	0.591	0.025	0.823	0.253	0.446	17.63	0.493	0.094
SAM3D	0.027	0.817	0.163	0.520	7.566	0.672	0.078	0.014	0.866	0.169	0.596	6.272	0.639	0.085
Ours	0.021	0.843	0.169	0.727	5.616	0.783	0.031	0.015	0.845	0.176	0.694	5.485	0.729	0.064



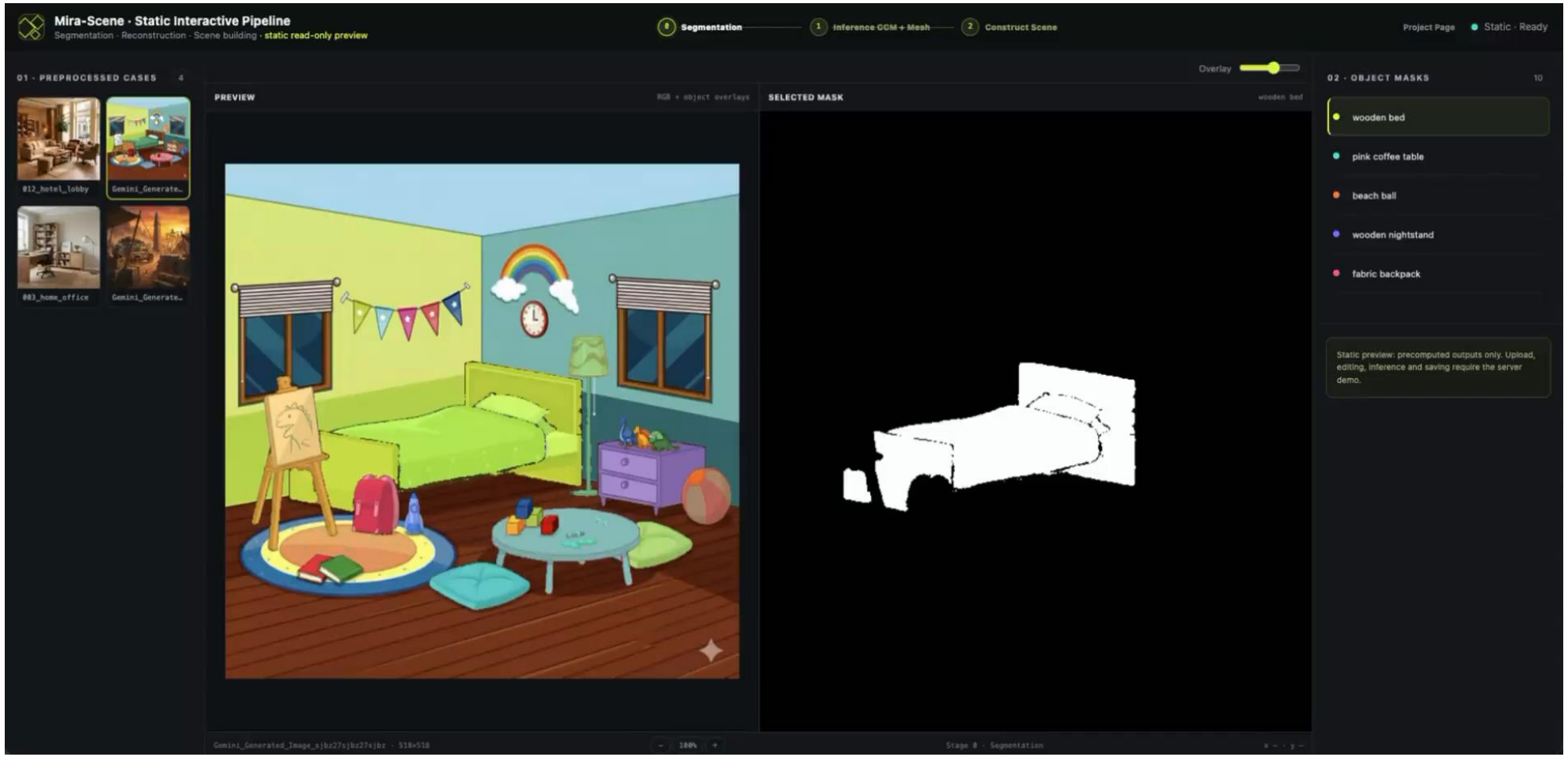
3D-first: Mira-Scene

- **A complete, ready-to-use pipeline**



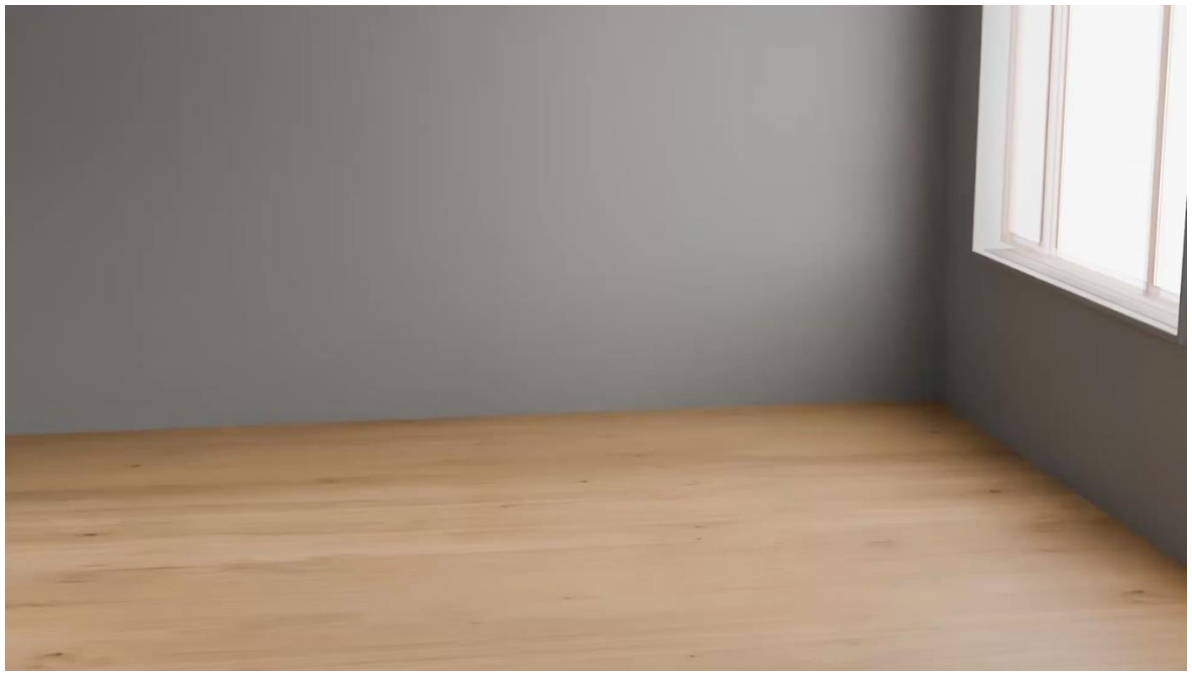
3D-first: Mira-Scene

- **A complete, ready-to-use pipeline**

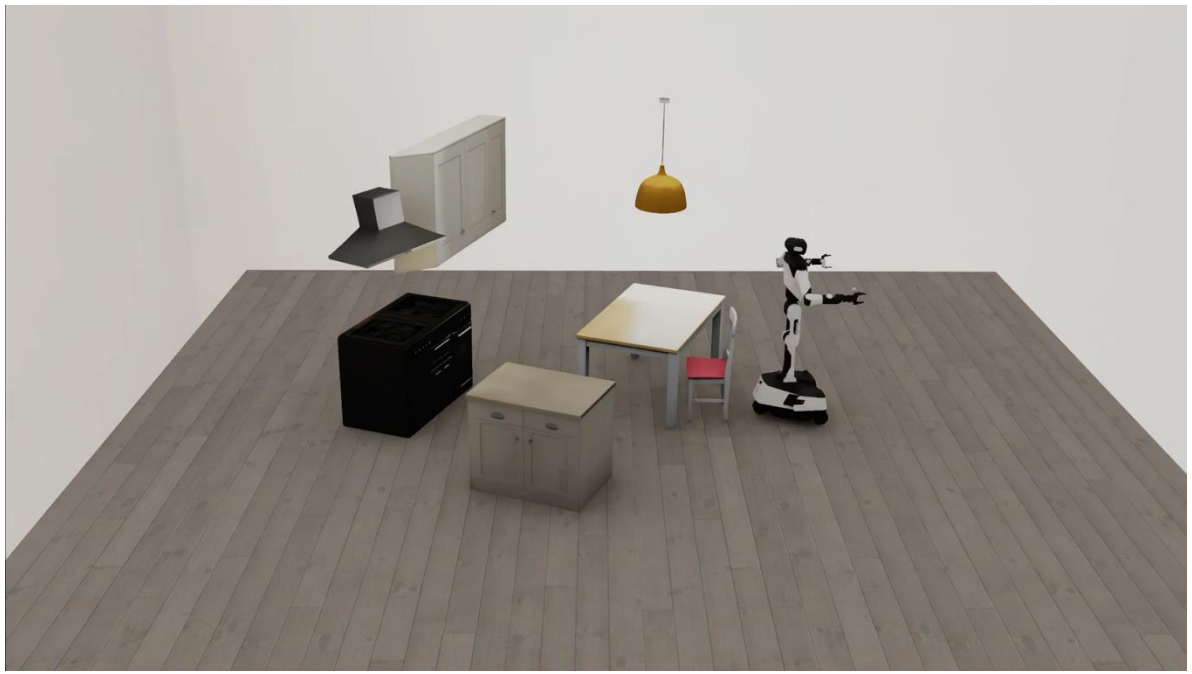


3D-first: Mira-Scene

- **Application**



Simulate, Edit in Blender



Robot Interaction in Isaac Sim

3D-first: Mira-Scene

- **Compared with Agentic Scene Generation, e.g. GPT6-Astra**

Condition Image

Mira-Scene + GPT-6 Astra

GPT-6 Astra From Scratch

For more comparison, please refer to the homepage: https://vast-eden.ai/?p=mira-scene

3D-first: Mira-Scene

- **Application in Embodied AI**

Tabletop scene building compared with GPT6-Astra

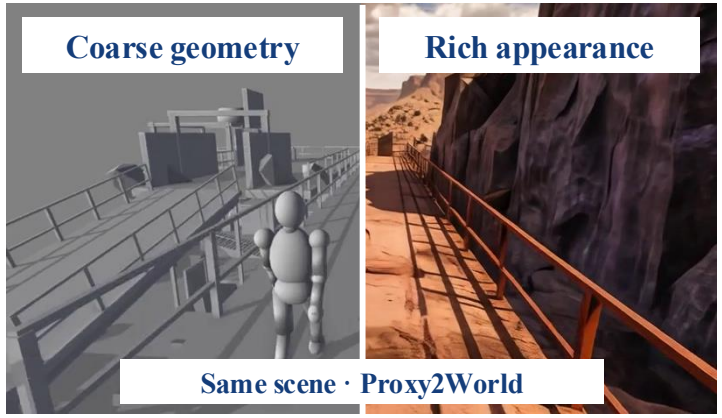
3D-first: Mira-Scene

- **Application in Embodied AI** *Tabletop scene + Robot interaction*

3D-first: Limits of Explicit Geometry

Three gaps between a 3D scene and a realistic world

01 Visual realism



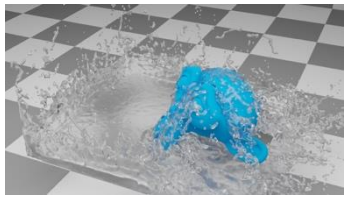
Geometry ≠ finished appearance

Detail, materials and lighting still need to be specified.

02 Complex physics

Cloth

Water



Folds · self-contact · splashes

Fine time steps and coupled solvers make realistic simulation costly.

03 Distant context

More than a local 3D mesh

Distant terrain, vegetation and sky need extra assets or representations.

Realistic worlds also require rich appearance, deformable dynamics and distant context.

Visuals: Proxy2World, Blender Manual and Greg Zaal / Poly Haven (CC0). Full sources in notes.

Bridging Two Worlds: 3D Control + Video Rendering

Editable 3D proxy

Layout · geometry · camera / motion

Generated appearance

Materials · lighting · visual detail

3D sets the constraints; video generation fills in appearance.

Proxy2World (2026) · official same-whitebox example; recorded generation, not live inference

Bridging Two Worlds: Proxy to Video

Toward Realistic, Persistent, and Interactive Worlds

Video-first

Looks great
Strong visual prior

State drifts
Long rollouts accumulate error

Actions brittle
Precise interaction is hard

3D-first

State explicit
Objects and poses are queryable

Physics-ready
Can attach colliders and dynamics

Assets limit
Rendering fidelity depends on assets

Proxy-to-Video

Looks + acts
Combines state and appearance

Proxy persists
Structure can be updated

Keep sync
Generated video must respect state

Video models learn how worlds look; explicit representations define what exists and how it changes.

Thanks for Listening !

孙阳天

CVMI Lab @ HKU

https://sunyangtian.github.io/